THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

Identification and characterisation of an altered gene in the novel ABA insensitive *Beyma* mutant of *Lotus japonicus*

Nur Fatihah Mohd Yusoff

BSc. MSc.

A thesis submitted for the degree of Doctor of Philosophy at

The University of Queensland in 2015

School of Agriculture and Food Sciences

Centre for Integrative Legume Research

Abstract

Various approaches can be implemented to identify a gene responsible for a phenotype of interest due to mutagenesis. Current next generation sequencing (NGS) technology allows whole genome sequencing of a mutant and accelerates the identification of mutation-induced polymorphisms in mutagenised organisms. In the model legume *Lotus japonicus* ecotype Miyakojima (MG-20), an abscisic acid (ABA) insensitive mutant called *Beyma* was previously isolated by ethyl methanesulphonate (EMS) mutagenesis and originally identified as a heterozygous dominant mutation. *Beyma* is slow-growing, wilty and incapable of regulating stomatal opening. A wild type segregant of the *Beyma* mutant (WTS) was also isolated from a self-generation of heterozygous *Beyma* mutants. ABA plays numerous roles in plant growth and development as well as morphogenetic responses including nodulation in legumes. Yet, there is a lack of ABA studies in legumes. Therefore, the *Beyma* mutant in *L. japonicus* allows a wide range of studies that will provide in-depth information of ABA signaling in nodulation as well as stress responses in legumes.

This project presents an attempt to identify a causal gene in the ABA insensitive *Beyma* using the NGS technology. Tissue from a homozygous *Beyma* mutant, WTS and MG-20 wild type (WT) plants was subjected to the whole genome sequencing, generating about 300 million paired end of short-sequence reads. The Kazusa MG-20 genome was used as reference for read mapping. Single nucleotide polymorphisms (SNPs) were called based on mutations in the *Beyma* and WTS genomes as compared to the re-sequenced MG-20 genome. As a preliminary study, three procedures of read mapping and variant calling were performed to undertake a genomic comparative analysis and identify the causal gene.

Sequencing of single genomes of the three plants showed a mutation occurred in every 208 kb (WTS) and 202 kb (*Beyma*) with a bias mutation of G/C-to-A/T changes at low percentage. Most mutations were intergenic. The mutation spectrum of the genomes was comparable in their individual chromosomes but each mutated genome has unique alterations, which are useful to identify causal mutations for their phenotypic changes. A total of 59 SNPs were identified as potential putative causal *Beyma* mutations, which were located in various annotated genes in the MG-20 genome. Verification of these mutations could not be done due to time constraint but will be performed in future. A candidate gene

approach was also carried out by selecting ABA-related genes based on their roles in ABA biosynthesis to signalling, directly or indirectly. Mutations were found in loci of *ABA INSENSITIVE* (*ABI*) *1*, *ABI2*, *HAB1*, *HAB2*, *ABI3*, *ABI4*, and *ABSCISIC ACID 8'-HYDROXYLASE 2* in both mutant genomes or only in the WTS genome. Unique mutations also occurred in *EARLY RESPONSIVE TO DEHYDRATION 7* and *ABSCISIC ACID 8'-HYDROXYLASE 1/ P450 CYP707A1* genes, which were predicted to be impaired in their downstream regions. Although the candidates were not affected in the essential region of the genes, the candidate gene approach has eliminated all the candidates as the putatively causal *Beyma* gene.

In order to intensify the identification of the causal *Beyma* gene, re-sequencing of the *Beyma* and WTS genomes was performed on pooled DNA. In this analysis, the presence of mutations was more frequent in both mutagenised genomes (~18-35% increase), resulting in higher rate of base changes and demonstrated that pooled DNA sequencing increased the mutation frequency. In addition, 69 unique *Beyma* SNPs were predicted to cause nonsynonymous changes and will be verified in future study. Nevertheless, a mutation (locus named chr3.CM0451.1060.r2.d) was found in both batches of sequencing. It was a C-to-T mutation, which changed glutamic acid to lysine in an F-box family gene. This gene could be the *Beyma* gene but it requires verification.

In conjunction with the genome sequence analysis, other analyses were also done to prepare plant materials for sequencing and future verification. Plants were subjected to ABA treatment on seed germination and root development to select good mutant lines and WTS plants. Outcross between *Beyma* and *L. japoncus* ecotype Gifu was also performed for the segregation analysis of the putative causal SNPs in the F2 plants carrying homozygous WT alleles. This project highlighted the overall molecular changes produced in the whole genome of MG-20 mutants due to EMS mutagenesis. In future, the identification of the causal *Beyma* gene will possibly show a novel gene involved in ABA sensitivity in legume systems. In addition, it should be of great interest for researchers in forward genetics in legume studies.

Declaration by author

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my research higher degree candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the policy and procedures of The University of Queensland, the thesis be made available for research and study in accordance with the Copyright Act 1968 unless a period of embargo has been approved by the Dean of the Graduate School.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis.

Publication during candidature

Peer-reviewed paper

Mohd-Yusoff NF, Ruperao P, Tomoyoshi NE, Edwards D, Gresshoff PM, Biswas B and Batley J (2014). Scanning ethyl methanesulphonate effects on the whole genome of *Lotus japonicus* using second generation sequencing analysis. *G3: GenesIGenomesIGenetics* 5: 559-567.

Conference abstract

Mohd Yusoff NF, Kazakoff S, Jacqueline B, Bandana B and Gresshoff P (Sept 2012). An altered gene in the novel ABA insensitive *Beyma* mutant of *Lotus japonicus*. Combio, Adelaide, Australia.

Publication included in this thesis

Mohd-Yusoff NF, Ruperao P, Tomoyoshi NE, Edwards D, Gresshoff PM, Biswas B and Batley J (2014). Scanning ethyl methanesulphonate effects on the whole genome of *Lotus japonicus* using second generation sequencing analysis. *G3: GenesIGenomesIGenetics* (submitted and accepted subject to minor revisions) – incorporated as Chapter 2.

Contributor	**Statement of contribution**
Nur Fatihah Mohd-Yusoff (Candidate)	Designed experiment (30%) Conducted experiment (55%) Wrote manuscript (100%)
Pradeep Ruperao	Conducted experiment (33%)
Nurain Emylia Tomoyoshi	Conducted experiment (1%)
David Edwards	Conducted experiment (1%)
Peter M Gresshoff	Designed experiment (15%) Edited paper (10%)
Bandana Biswas	Designed experiment (25%) Conducted experiment (10%) Edited paper (45%)
Jacqueline Batley	Designed experiment (30%) Edited paper (45%)

Contributions by others to the thesis

Chapter 4

Dr Stephen Kazakoff and Mr. Pradeep Ruperao conducted experiments and contributed to experimental design of read mapping and SNP calling analyses.

Chapter 5

Dongxue Li and Dr Satomi Hayashi conducted experiments for re-sequencing. Ms. Jenny Lee conducted read mapping and SNP calling analyses.

All Chapters

Supervisors, Prof. Peter M Gresshoff, Dr Bandana Biswas and Prof. Jacqueline Batley, assisted in experimental design and editing of the writing.

Statement of parts of the thesis submitted to qualify for the award of another degree

None

Acknowledgements

First, I would like to thank Allah for giving me strength and wisdom to complete this study. It is my pleasure to acknowledge the Ministry of Education, Malaysia and Universiti Putra Malaysia for sponsoring my study at the University of Queensland. I would also like to thank the Centre for Integrative Legume Research (CILR) and School of Agriculture and Food Sciences for giving the opportunity to conduct my PhD research and providing funds for conference attendance and travel.

No words will be able to describe my heartfelt gratitude and appreciation to my supervisor Prof. Peter M Gresshoff for his constant guidance, invaluable advice, stimulating discussions and ideas throughout the course of this project. Special thanks are extended to other advisors, Dr Bandana Biswas and Prof. Jacqueline Batley, for their advices, comments and guidance whenever sought. I really appreciate their patience and understanding.

I would also like to thank all CILR staff and students, especially former CILR student Dr Stephen Kazakoff, for their technical assistance whenever needed, the brainstorming discussion we had together and creating a good environment working in the laboratory. My appreciation also goes to all of my friends, who studied in the University of Queensland, for their friendships and supports while living in Brisbane.

Not forgotten, my special thanks go to Ayah, my siblings and family for giving me support while studying oversea. Last but not least, I would like to express my gratitude to my beloved husband, Uzaeir, and kids, Uthman and Ulfah, for always being there and supporting my PhD. I also appreciate my husband for his technical assistance in my PhD research whenever he visited me in Brisbane.

Keywords

legumes, *Lotus japonicus*, next generation sequencing, *Beyma*, EMS mutagenesis, abscisic acid, candidate gene

Australian and New Zealand Standard Research Classifications (ANZSRC)

ANZSRC code: 060702, Plant Cell and Molecular Biology, 20%

ANZSRC code: 060705, Plant Physiology, 20%

ANZSRC code: 060102, Bioinformatics, 60%

Field of Research (FoR) Classification

FoR code: 0604, Genetics (35%)

FoR code: 0607, Plant Biology (65%)

Table of Contents

List of figures

List of tables

Chapter 5

List of abbreviations

AB	Alternate bases
ABA	Abscisic acid
ABA-GE	Abscisic acid glucose ester
ABF	ABA responsive element-binding factor
ABI	Abscisic acid insensitive
BG1	β-glucosidase
BWA	Burrows Wheeler Aligner
CCaMK	Calcium/ calmodulin-dependent protein kinase
EMS	Ethyl methanesulphonate
ENF	Enhanced nitrogen fixation
ERA	Enhanced Response to ABA
ERD	Early responsive to dehydration
EST	Expressed sequence tags
GA	Genome Analyzer
GORK	Gated outwardly rectifying K+ channel
GRF	Growth regulating factor
HAB	Homology to ABI
KAT	Potassium channel in *A. thaliana*
LATD	Lateral root organ defective
LRR	Leucine rich repeat
MG-20	Miyakojima
NGS	Next generation sequencing
OST	Open stomata
PE	Paired end reads
PP2C	Type 2 protein phosphatase
PYR/PYL/RCAR	Pyrabactin resistance/ pyrabactin-like/regulatory component of ABA receptor
RBOHD	Respiratory burst oxidase homologue
RPK	Receptor protein kinase
SGS	Second generation sequencing
SLAC	Slow anion channel associated
SLAH	SLAC homologue
SNPs	Single nucleotide polymorphisms

SnRK	Serine/threonine protein kinase
SOLiD	Sequencing by Oligo Ligation Detection
SSR	Simple sequence repeats
STA	Sensitive to ABA
TAC	Transformation-competent artificial chromosome
TILLING	Targeting induced local lesions in genomes
VP	Viviporous
WT	Wild type
WTS	Wild type segregant of the *Beyma* mutant

Chapter 1
General introduction

1.1 Abstract

Model plants are adopted in many researches to provide knowledge that is beneficial to the improvement of crop quality and yield production. The model legume *L. japonicus* (Japanese trefoil) has been utilised in a wide range of physiological and molecular biology analyses including genome sequence project. At present, plenty of information on *L. japonicus* is easily accessible that facilitated numerous research projects on legumes. The genomic sequence of *L. japonicus* ecotype Miyakojima and other legumes, such as *Glycine max* (soy bean) and *Medicago truncatula* (barrel medic) is also available publicly, offering a good platform in the development of legume research. With the current technology, advanced analyses such as next generation sequencing have been developed to explore the biological secrets of legume plants including hormone responses, development and diseases. Many genes have been identified playing crucial roles in the action of plant hormones such as ABA. This introductory chapter summarises the current knowledge, which is known relating to legumes, sequencing analyses and ABA.

1.2 Introduction

One of the most important groups in the crop plantation world is members of Fabaceae (or Leguminosae) family called legumes. Legumes, such as soybean, bean, pea and peanut, are highly important food, feed and biofuel crops (Ferguson et al., 2010). Various analyses have been performed to enhance the knowledge of the legume system and to contribute ideas for the development of agricultural and environmental purposes. Thus, model plants, namely *L. japonicus* and *M. truncatula,* were introduced to achieve these objectives.

L. japonicus is a diploid legume (2n=12), has determinate type of nodulation, a short life cycle (2 – 3 months) and a relatively small genome size which is suitable for molecular and genetic analyses (Handberg and Stougaard, 1992; Lohar et al., 2001; Sato et al., 2008). In legumes, the phytohormone ABA has been shown to play roles as a negative regulator (Bano and Harper, 2002; Ferguson and Mathesius, 2003; Suzuki et al., 2004) in which it inhibits nodulation and bacterial infection (Ding et al., 2008). However, most of the work linking ABA to nodulation has relied on the application of exogenous hormones or inhibitors, measuring concentration changes in large tissue samples in response to various developmental steps and comparisons between WT and nodulation-control mutants (Fujita et al., 2006; De Smet et al., 2006).

Biswas et al. (2009) isolated an ABA insensitive mutant, termed *Beyma*, in *L. japonicus* ecotype MG-20 using EMS mutagenesis. The stable *Beyma* mutant has wilty phenotype and slower growth than its WT. Analysis of the stable *Beyma* mutant also showed that ABA inhibition is local and not involved directly in systematic autoregulation of nodulation (Biswas et al., 2009). Thus, further analysis of this mutant will improve understanding of the ABA-inhibition of nodulation in *L. japonicus* system. ABA is crucial for plant responses to environmental stimuli, such as drought, cold and salinity. Although a lot is known about ABA responses in plants, mainly through studies on *Arabidopsis thaliana*, very little of that involves studies on legumes. Therefore, the *Beyma* mutant in *L. japonicus* allows wide ranges of studies that will provide in-depth information of ABA signaling in nodulation as well as stress responses in legumes.

Next generation sequencing (NGS) technologies have been widely applied in forward genetics and genomics studies. At present, the cost of sequencing the whole genome has reduced and various NGS tools have been developed, offering opportunities to undertake NGS approaches in digging the genetic makeup of individuals in plant research. The NGS technologies are also anticipated to contribute the development of crop breeding program (Varshney et al., 2009, 2014; Edwards et al., 2013). In addition, the establishment of MG-20 genome sequence by the research of the Sato group (2008; at the Japanese Kazusa DNA Research Institute) allowed the application of NGS in our project to identify a putatively causal gene in ABA insensitive *Beyma* mutant.

This project highlights the overall molecular changes produced in the whole genome of *L. japonicus* mutants due to EMS mutagenesis. Furthermore, the identification of the causal *Beyma* mutations showed potential genes that are involved in ABA mechanisms and allowed the assessment of the noncausal mutations which resulted from the original mutagenesis. This survey of collateral damages has been hitherto overlooked in genome-phenome linkages. Information from this study will also enhance understanding of ABA signaling at the molecular level in legumes and may be applicable to the improvement of legume commercial production.

1.2.1 *Lotus japonicus*

For many years, model plants such as *A. thaliana* and *Oryza sativa* (rice) have been used as models in dicotyledon and monocotyledon systems, respectively, to gain various fundamental knowledge and information in plant biology. However, they do not nodulate and therefore, cannot be used to study some crucial aspects in legume systems such as symbiotic nitrogen fixation and legume breeding programmes. Hence, two leguminous plants, *L. japonicus* and *M. truncatula* have been introduced and adopted as model plants to reveal insight into the leguminous systems for agronomic benefits (Handberg and Stougaard, 1992; Udvardi, 2001; Stacey et al., 2006; Sato and Tabata, 2005). In this study, we adopted *L. japonicus* as our model plant.

L. japonicus is indigenous to the Far East (including regions of Japan, China and Korea) and a member of Loteae subfamily under the same family (Papilionoideae) with other legumes, *G. max* and *M. truncatula* (Udvardi et al. 2005; Melchiorre et al., 2009). *L.*

japonicus has relatively small genome (~470 Mb) with six chromosomes (2n=12), short generation time (3-4 months), indeterminate flowering, straight seed pods and a large number of small seeds per plant (Figure 1.1). *L. japonicus* also has perennial growth, is able to self-fertilise and susceptible to transformation by *Agrobacterium tumefaciens.* These characteristics provide great advantages in various studies such as molecular genetics and functional genomics (Handberg and Stougaard, 1992; Jiang and Greeshoff, 1997; Szczyglowski and Stougaard, 2008). In symbiotic nitrogen fixation, *L. japonicus* roots interact with their symbiont, *Mesorhizobium loti* to produce determinate spherical nodules (Figure 1.1) opposite to *M. truncatula* that develops indeterminate nodulation (Hayashi et al, 2000; Saeki and Kouchi, 2000; Sato and Tabata, 2005; Udvardi et al., 2005; Ferguson et al., 2010).

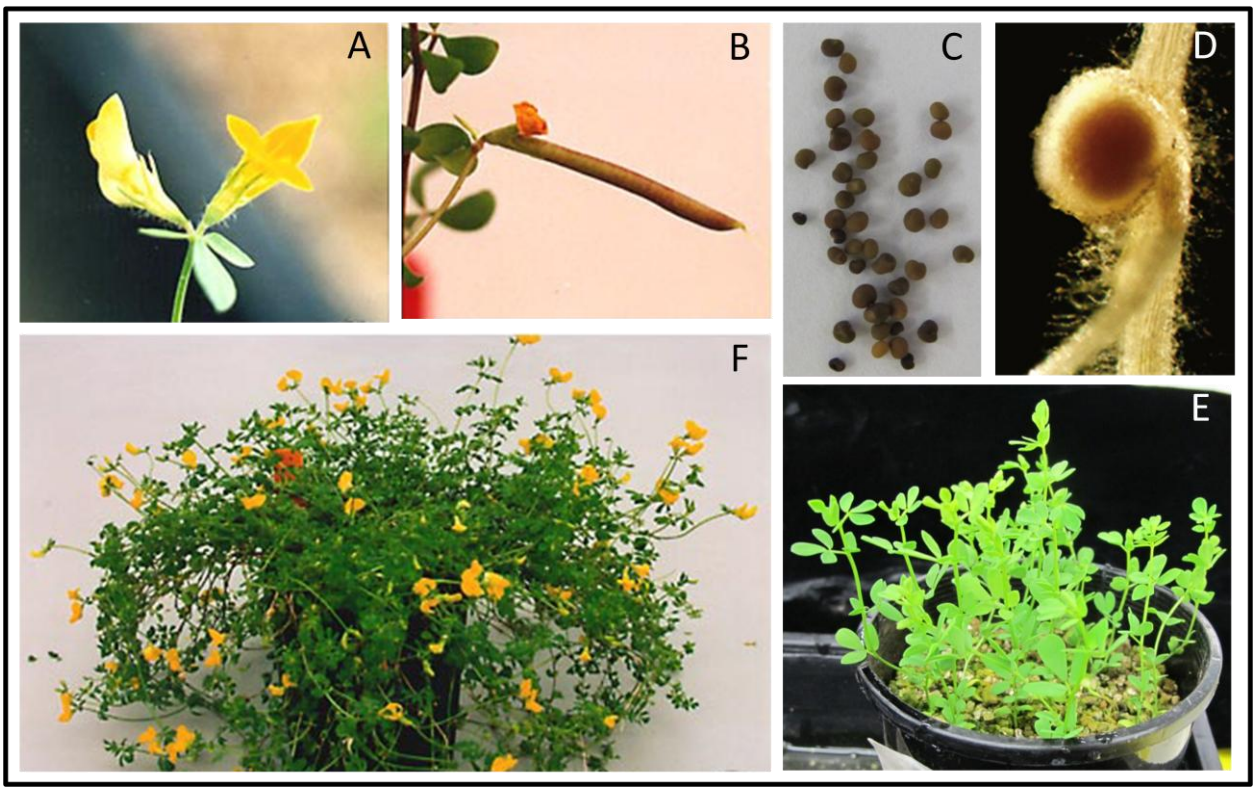

Figure 1.1: Phenotypic characteristics of *L. japonicus*. **A**. A mature flower (8-12 mm in length); **B**. a mature seed pod (about 3 cm in length); **C**. harvested seeds (2-4 mm in diameter); **D**. a root nodule (about 1 mm in diameter); **E**. one-month old seedlings and **F**. mature plant exhibiting abundant flowering.
Adapted from; Jiang and Gresshoff (1997); Szczyglowski and Stougaard (2008).

1.2.2 *Lotus japonicus* genome project

In order to investigate deeper in genetics of other legumes relative to *L. japonicus,* acquisition of genomic DNA sequence of *L. japonicus* will provide great benefit to legume molecular biology. The genomic sequencing of MG-20 was initiated by Sato et al. (2001) using transformation-competent artificial chromosome (TAC) cloning. Obtained sequences were utilised to generate DNA markers for linkage mapping on an F2 population of a cross between *L. japonicus* ecotype Gifu and MG-20, developed by Hayashi et al. (2001). Subsequently, structural analyses of the *L. japonicus* genome were performed in a few stages in order to comprehensively analyse the TAC clone libraries and organization of putative genes that resulted from the sequenced clones (Nakamura et al., 2002; Kaneko et al., 2003; Asamizu et al., 2003; Kato et al., 2003). The whole-genome sequence of MG-20 was then successfully constructed covering a total length of 315,073,275 bp sequence (67 % of the 472-Mb genome) and 91.3 % of gene space is located in the determined genome (Sato et al., 2008). Currently, Sato and Andersen (2014) announced that the latest version of the MG-20 genome was successfully determined using NGS technology, covering ~87 % of the total genome length.

1.2.3 Bioinformatics resources available on *Lotus japonicus* and other legumes

At present, researches are able to acquire numerous publicly available bioinformatics materials of legumes for computational biology and comparative analyses like functional and structural genomics. In *L. japonicus*, such information and material are provided by many resources, as summarised by Sato and Tabata (2005; Table 1.1). The website of the Kazusa DNA Research Institute (http://www.kazusa.or.jp/lotus/index.html) contains sequencing data on six chromosomes of *L. japonicus* (Sato et al., 2008) and tools for searching of annotated protein-encoding genes. Each protein-encoding gene has been pre-searched for its homologous sequences in *A. thaliana, M. truncatula* and/or soybean, facilitating the gene analysis.

A total of 788 simple sequence repeats (SSR) and 80 derived cleaved amplified polymorphic sequences used for linkage mapping are also available at http://www.kazusa.or.jp/lotus/clonelist.html (Sato et al., 2008). Meanwhile, a database on

the expressed sequence tags (EST) and tentative consensus of *L. japonicus* can be obtained at *Lotus japonicus* EST Index (http://est.kazusa.or.jp/en/plant/lotus/EST/) for transcriptome analysis. Resources service for *L. japonicus* and soybean biology materials such as seeds, phenotypic information and clones is also available at the website of Legume Base (http://www.legumebase.brc.miyazaki-u.ac.jp/). Interestingly, Kato et al. (2000) have sequenced and assembled the chloroplast genome of *L. japonicus* (150,519 bp) which is available at http://www.kazusa.or.jp/lotus/Cp/index.html. Meanwhile, the mitochondrial genomic sequence of *L. japonicus* has been analysed by Kazakoff et al. (2012), resulting in the sequence assembly of 380,861-bp in length.

Table 1.1: Bioinformatics resources available on *L. japonicus* and other legumes.

Database name	**Information and materials**	**Resource**
Miyakogusa	*Lotus genome*	http://www.kazusa.or.jp/lotus/index.html
Lotus japonicus EST Index	*Lotus* EST	http://est.kazusa.or.jp/en/plant/lotus/EST/
L. japonicus Gene Index	*Lotus* EST and transcript	http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/gimain.pl?gudb=l_japonicus
Legume Base	*Lotus* and soybean (seeds, phenotype, clones, etc.)	http://www.legumebase.brc.miyazaki-u.ac.jp/
TAGdb	*Lotus* pair-reads *Pongamia* pair-reads	http://flora.acpfg.com.au/tagdb/
Reverse Genetics UK	*Lotus* TILLING facility *Medicago* TILLING facility	http://revgenuk.jic.ac.uk/TILLING.htm
Phytozome	Soybean genome *Medicago* genome	http://www.phytozome.net/
Soybean	Soybean genetic map Soybean synteny browser	http://soybase.org/
Noble Foundation	*Medicago* expression atlas	http://bioinfo.noble.org/gene-atlas/v2/
Legume IP	Legumes atlas	http://plantgrn.noble.org/LegumeIP/

In 2009, Biswas et al. isolated an ABA insensitive mutant called *Beyma* from MG-20. The mutant together with its WT and WTS of the mutant were deep-sequenced by Prof. Dave Edwards group from the Australian Centre for Plant Functional Genomics

(ACPFG, Brisbane, Queensland) using pair-read sequencing technology (data submitted for publication). All the 100 bp paired reads generated were uploaded in the database, TAGdb (http://flora.acpfg.com.au/tagdb/), which can be used for the identification of mutation. The TAGdb database also provides sequence databases on other plants such as Pongamia and *Brassica rapa*. All of these resources facilitate the computational analyses that are required not only for *L. japonicus* research but also for other legumes. On the other hand, Phytozome (http://www.phytozome.net/) is another useful database that compiles information on various plants including legume plants such as soybean, *M. truncatula* and *Pisum sativum*. This database contains annotated genes of the plant genomes and is capable of identifying their high sequence identity and microsynteny between the plants. Some other resources available on legumes are also listed in Table 1.1.

1.2.4 Next generation sequencing

The effort of determining DNA sequence in living organisms commenced in the past few decades. Single-stranded bacteriophage DNA was the first full genome to be sequenced using a “plus and minus” method (Sanger et al., 1977a). After that, a chain-terminating inhibitors technique was introduced (Sanger et al., 1977b), producing more accurate DNA sequences rapidly. Sanger sequencing techniques have been applied to determine genome sequences of various plant species such as *A. thaliana* (the first complete plant genome; The Arabidopsis Genome Initiative, 2000), *L. japonicus* (Sato et al., 2008) and soybean (Schmutz et al., 2010). This technology is known as “first generation sequencing”. With a demand in obtaining high throughput data of genetic information, the DNA sequencing technologies continued to develop. Consequently, NGS or second generation sequencing (SGS) was introduced involving different principles of sequencing, such as sequencing-by-synthesis, oligonucleotide probe ligation and pyrosequencing (Metzker 2010; Pareek et al., 2011; Liu et al., 2012; Thudi et al., 2012).

Development of NGS technologies is conquered by a number of companies, which are continuously improving the quality of sequencing platforms. In 2005, the first commercial sequencing machine, GS 20, was developed by a company called 454 Life Sciences, which was then taken over by Roche Applied Science. Later, several other platforms were launched by Roche, adopting pyrosequencing (shotgun sequencing

procedure) mechanism based on the detection of pyrophosphate released during nucleotide incorporation (Pareek et al., 2011). Roche 454 initially generated 100-150 bp reads before upgrading to GS FLX, which is able to produce longer reads of up to 700 bp in length with high accuracy (Liu et al., 2011).

In 2006, the Genome Analyzer (GA) platform was developed by Solexa company before being purchased by Illumina in the following year. The sequencer adopts sequencing by synthesis principle with reversible terminators. The read length of GA increased from 75 bp to 150 bp paired end with the improvement of GA systems (Liu et al., 2012). Illumina then launched HiSeq 2000 and a bench top sequencer MiSeq, which adopt the same principle as GA and could sequence paired end reads of up to 150 bp in length. These three systems are able to produce a significant yield of bases greater than quality of 30 and require the same range of DNA amount as template, which is 50-1000 ng (Liu et al., 2012; Quail et al., 2012). However, MiSeq requires a shorter time (~27 hours) to run its workflow including cluster generation (Quail et al., 2012).

Another company that also develops sequencing instruments was Agencourt who launched Sequencing by Oligo Ligation Detection (SOLiD). In 2006, Applied Biosystems purchased Agencourt and released new instruments of SOLiD. These sequencers adopt the technology of two-base sequencing based on oligonucleotide probe ligation. Read length of SOLiD sequencers was shorter than Roche 454 and Illumina sequencers, in which it was initially 35 bp and later, was improved to 50 bp (Thudi et al., 2012). As reviewed in Liu et al. (2012), HiSeq 2000 provides cheaper cost than applying 454 and SOLiD technologies. However, these NGS instruments have advantage of producing high throughput output with lower cost as compared to Sanger technology. In addition, other technologies such as Ion Torrent Personal Genome Machine (Life Technologies) and Pacific Biosciences RS (Pacific Biosciences) are currently available in market (Quail et al., 2012; Thudi et al., 2012). The latest sequencing technology called third generation sequencing has been developed, generating higher accuracy of reads with longer sequence in more rapid and higher throughput fashions (Quail et al., 2012; Thudi et al., 2012; Koren et al., 2013).

1.2.5 Next generation sequencing in the legume genomes

Since legumes are one of the most important crops in the world, essential knowledge or information was anticipated to contribute to the development of crop yield. Genome projects of two model legumes, *M. truncatula* and *L. japonicus*, and soybean were extensively conducted, leading to the advances of genomics and genetics research in legumes. The genome of *M. truncatula* and *L. japonicus* was sequenced using the same strategy, in which sequence genespaces were favourably determined in the genomes or clone by clone strategy (Cannon et al., 2005; Young et al., 2005; Sato et al., 2007). Contrary to soybean, its genome was constructed using shotgun sequencing (Schmutz et al., 2010). The latest report on legume genome sequencing was on *Phaseolus vulgaris* L. (common bean), which adopted whole genome shotgun sequencing (Schmutz et al., 2014).

At present, NGS technologies have been widely applied in the legume genome projects to obtain high throughput data rapidly. Reduced cost in adopting NGS tools has also raised the implementation of the NGS technologies (Pareek et al., 2011; Liu et al., 2012; Thudi et al., 2012). In legumes, whole genome sequencing of *Pongamia pinnata* was performed in 2010 (Peter Gresshoff, personal communication) using Illumina GA, which led to the assembly of its organellar genomes of chloroplast and mitochondrion (Kazakoff et al., 2012). In addition, Sato and Andersen (2014) constructed a new version of *L. japonicus* assembled genome sequence, which was determined using Roche 454 GS FLX and Illumina platforms. Output from these NGS sequencers were assembled with longer sequence libraries obtained from the clone by clone approach. Similar platforms were also adopted by Schmutz et al. (2014) in generating a high quality reference genome of common bean. Integrating Roche and Illumina platforms, the sequence assembly of common bean genome was organised into eleven chromosome-scale pseudomolecules. In order to enhance the development of legume researches, sequence databases can be retrieved publicly online as listed in Table 1.1. Following to these, information on genome sequences and gene annotation of legumes can be used to perform comparative genomics and transcriptomic analyses between legume plants.

1.2.6 ABA perception and signaling

Plants have several classes of hormones that are pleiotropic in their effects in growth and morphogenetic responses. One of these classes is ABA, a small lipophilic sesquiterpenoid (C_{15}). It is suggested to be synthesised either in a direct pathway from farnesyl pyrophosphate or indirect pathway by cleavage of a carotenoid C_5 precursor, isopentenyl diphosphate (Cutler and Krochko, 1999). ABA glucose ester (ABA-GE), an inactive conjugate, has been postulated to play an essential role in long-distance signaling of ABA. ABA-GE is synthesised in the cytosol but has low permeability in plasma membrane. ABA-GE transporter assists the ABA-GE to move into apoplastic pathway for translocation from the root to shoot in the xylem and from the shoot to root in the phloem (Jiang and Hartung, 2008; Wasilewska et al., 2008). The ATP-binding cassette transporter family has been postulated to be involved in transporting ABA into apoplastic space in limited rate (Jiang and Hartung, 2008; Umezawa et al., 2010). Cleavage of glucose from ABA is catalised by apoplastic and endoplasmic reticulum β-glucosidase (BG1; Figure 1.2). The same level of free active ABA was found in both *Atbg1* mutant and WT, but the mutant was lacking of free ABA upon dehydration stress. The results indicated that the BG1 activity increases the content of active free ABA in extracellular level and induces intracellular ABA signaling under both normal and stress conditions (Lee et al., 2006).

A network module of ABA perception and signaling have been summarised in numerous reviews, showing the interaction between ABA with its receptors and regulatory networks. Several plasma membrane and intracellular ABA receptors were reported to be involved in ABA perception, including pyrabactin resistance (PYR)/ pyrabactin-like (PYL)/ regulatory component of ABA receptor (RCAR), Flowering Time Control Protein A, magnesium chelatase and G-proteins (Cutler et al., 2010; Raghavendra et al., 2010). An early ABA signaling engages the perception of ABA by a nucleocytoplasmic PYR/PYL/RCAR complex which possesses gate (proline cap) and latch (leucine lock) loops as ABA-binding pocket. The gate-ABA-latch binding leads to the reconformation of PYR/PYL/RCAR proteins and the anchorage of type 2 protein phosphatase (PP2C) such as *ABI1* and *ABI2* (Hubbard et al., 2010; Cutler et al., 2010; Umezawa et al., 2010). The interaction inhibits PP2C activity which subsequently induces the phosphorylation and activation of serine/threonine protein kinase (SnRK2) such as *OPEN STOMATA 1* (*OST1*). This leads to the activation of ion channel genes (*SLOW ANION CHANNEL ASSOCIATED*

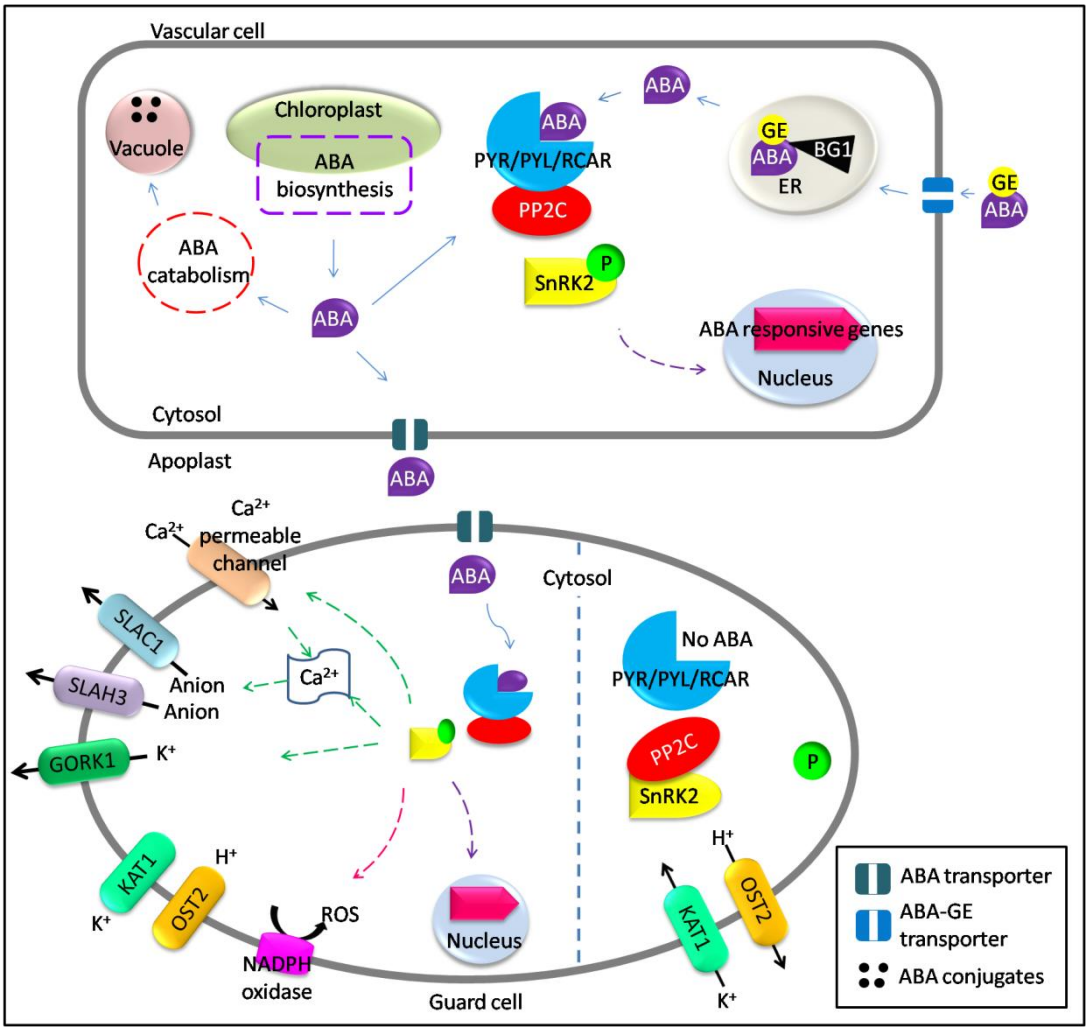

Figure 1.2: ABA perception and signaling. ABA biosynthesis occurs in chloroplast except the last two which are cytoplasmic. ABA is catabolised putatively in cytosol producing ABA conjugates that accumulate in vacuole of vascular cells. Apoplastic ABA-GE moves into vascular cell via an unknown ABA-GE transporter and is cleaved by endoplasmic reticulum (ER) BG1, releasing free ABA. Cytoplasmic PYR/PYL/RCAR complex binds to the ABA and anchors PP2C, leading to phosphorylation of SnRK2 which promotes the expression of nucleic ABA responsive genes. In guard cell, SnRK2 also elevates cytosol Ca^{2+} content and activates plasma membrane NADPH oxidase that releases secondary messengers such as ROS. Ca^{2+}-dependent and -independent signals deactivate ion channel *POTASSIUM CHANNEL IN A. THALIANA 1* (*KAT1*) and *OST2* genes but activate other ion channels (*SLAC1*, *SLAH3* and *GORK1*) for anion and K^+ effluxes, which leads to stomata closure. In the absence of ABA, PP2C inhibits the phosphorylation of SnRK2 which induces K^+ influx and H^+ efflux by ion channel *KAT1* and *OST2*, respectively, leading to stomata opening. Solid arrows represent biosynthesis and catabolism pathways. Dash arrows represent ABA signaling pathway. Adapted from; Cutler et al. (2010), Kim et al. (2010), Joshi-Saha et al. (2011), Hauser et al. (2011).

1, SLAC1; *SLAC1 HOMOLOGUE 3, SLAH3* and *GATED OUTWARDLY RECTIFYING K+ CHANNEL 1*, *GORK1*) and nucleic ABA-responsive genes such as leucine zipper transcription factors (*ABA RESPONSIVE ELEMENT-BINDING FACTOR 2*, *ABF2*/ *AREB1* and *ABI5*) and plasma membrane NADPH oxidases (*RESPIRATORY BURST OXIDASE HOMOLOGUE*, *AtRBOHD* and *AtRBOHF*), resulting in the ABA-inhibition of stomatal opening through which a plant responses to environmental cues such as drought (Cutler et al., 2010; Umezawa et al., 2010; Hauser et al., 2011).

A summary of ABA perception and signaling is shown in Figure 1.2, demonstrating the function of PP2C and SnRK2 as negative and positive regulators, respectively. In guard cells, ABA elevates cytosol Ca^{2+} by inducing the influx of Ca^{2+} into cytosol. Both of Ca^{2+}-dependent and Ca^{2+}-independent pathways trigger a signal to activate the efflux of anions by *SLAH3* and *SLAC1* as well as inactivate the influx of K^+ (by *KAT1*) and the efflux of H^+ by *OST2*. These cause depolarization of the plasma membrane and induce the release of K^+ by *GORK1* from guard cells, which leads to reduction of turgor and volume of the guard cells and therefore, closing of stomata (Schroeder, 1992; Sirichandra et al., 2009; Kim et al., 2010; Joshi-Saha et al., 2011).

1.2.7 ABA roles and genes involved in legumes

In *A. thaliana*, *a*uxin transport and signaling promote the initiation of lateral root primordia. Whilst ABA regulates the lateral root development after the emergence directly or indirectly (De Smet et al., 2006). ABA induces *ENHANCED RESPONSE TO ABA 1* (*ERA1*) in order to repress activity of auxin-induced *ABI3*/ *VIVIPOROUS* (*VP1*) during auxin signaling for lateral root development (De Smet et al., 2006). ABA also limits penetration and growth of adventitious roots which are gibberellin-promoted and ethylene-induced at different levels in rice (Steffens et al., 2006). In legume plants, ABA regulation in root development is different probably due to the capability of legume roots in producing nodules. In *M. truncatula*, *LATERAL ROOT ORGAN DEFECTIVE* (*LATD*) encodes NRT1 (PTR) transporter protein which functions in development of lateral and primary roots as well as nodules. This gene is regulated by ABA which could restore the root meristem defects of *latd* mutant. The presence of 10 µM exogenous ABA was sufficient to induce the formation of meristem cells in primary and lateral root tips. The *latd* mutant also exhibited a

decreased sensitivity to ABA in both stomatal closure and seed germination (Liang et al., 2007; Yendrek et al., 2010), suggesting differential regulation of ABA in root development.

ABA is also known as a negative regulator of nodulation by controlling processes required for nodule development, including bacterial infection, Nod factor signaling and consequently, a nodulin gene expression (Ferguson and Manthesius, 2003). s*ensitive to ABA* (*sta-1*) mutant showed hypersensitivity to ABA in nodulation of *M. truncatula* but its ABA sensitivity was reduced in Nod factor signaling. These results showed that the *STA-1* role is crucial in the initial stage of Nod factor signaling (Ding et al., 2008). Endogenous ABA also controls nodule number and activity of nitrogen fixation by reducing production of nitric oxide in nodules. A *L. japonicus* mutant, *enhanced nitrogen fixation* (*enf1*) showed lower ABA content and ABA sensitivity which in turn increased nodule number and enhanced nitrogen fixation activity of *enf1* mutant (Tominaga et al., 2010).

1.2.8 Description of *Beyma*

Beyma is the ABA insensitive mutant that was isolated from EMS-treated MG-20 by Biswas et al. (2009). The *Beyma* mutant develops slower than its WT, producing smaller leaves, a shorter shoot and reduced number of lateral roots as well as a wilty phenotype. *Beyma* is not mutated in its nodulation since the development of bacteroid-infected nodules occurred upon *M. loti* inoculation although the produced nodules were slightly smaller. However, *Beyma* produced more nodules than WT upon application of exogenous ABA, indicating ABA-inhibition of nodulation of *Beyma* was impaired. Analysis of the stable *Beyma* mutant also showed that ABA inhibition is local and not involved directly in systematic AON.

An ABA sensitivity test on root growth of *Beyma* identified a reduced ABA sensitivity with 3:1 ratio, indicating the *Beyma* gene segregates as a dominant mutation with monogenic trait. Seed germination of *Beyma* also showed a decrease in ABA sensitivity. In addition, *Beyma* was unable to regulate its ABA- and drought-mediated stomatal closure upon ABA treatment and drought stress. These treatments resulted in reduced number of closed stomata and increased dryness susceptibility in the presence of ABA and drought stress, respectively (Biswas *et al.*, 2009). A highly similar phenotype to the *Beyma* mutant was found in *A. thaliana*, *abi1* mutant (Merlo et al., 2001). However, the ortholog of *L.*

japonicus to *AtABI1* was not altered in both MG-20 WT and *Beyma* (Biswas et al., 2009). Fujii and Zhu (2009) also reported similar phenotype on triple mutant of protein kinases (*OST1*, *SnRK2*.2 and *SnRK2*.3) in *A. thaliana*, in which ABA-inhibition of seed germination was reduced and susceptibility of dehydration was increased. These results suggested that *Beyma* might be defective in its ABA-related signaling. Therefore, further analysis of this mutant will improve understanding in ABA signaling and ABA-inhibition of nodulation in *L. japonicus* system.

1.3 Statement of thesis aims and structures

This project aimed to identify the mutated gene responsible for the ABA insensitive phenotype in *Beyma*. Investigation of the candidate sequence of ABA genes will facilitate the identification of the putative causal mutation of *Beyma*. In addition, adopting NGS technology will show EMS effects on the mutant genomes and subsequently, help to identify the causal gene. Thus, two approaches were undertaken; firstly was a candidate gene approach and secondly, SGS method. This thesis was divided into four chapters representing the project progress to achieve the main objective. Chapter 2 was to identify SNPs and show the EMS effects on nucleotide sequences in the EMS mutagenised MG-20 genomes. Chapter 3 represents a candidate gene approach with aimed to identify a mutation uniquely in selected candidate genes which are related to ABA signaling and guard cell signaling transductions. The SGS method was continued (Chapter 4 and 5) to list all putative mutated sequences in WTS and *Beyma* mutant for the identification of causal SNPs in ABA insensitive *Beyma*.

Chapter 2

Scanning ethyl methanesulphonate effects on the whole genome of *Lotus japonicus* using second generation sequencing analysis

Preface:

This chapter shows base alterations occurring in the whole genome due to mutagenesis, and has been published in *Genes*|*Genomes*|*Genetics* (2015, Vol. 5, pp. 559-567).

Scanning the Effects of Ethyl Methanesulfonate on the Whole Genome of *Lotus japonicus* Using Second-Generation Sequencing Analysis

Nur Fatihah Mohd-Yusoff,*,† Pradeep Ruperao,‡,§ Nurain Emylia Tomoyoshi,* David Edwards,‡, Peter M. Gresshoff,*,1 Bandana Biswas,* and Jacqueline Batley*,****

*Centre for Integrative Legume Research, School of Agriculture and Food Science, The University of Queensland, St Lucia, Brisbane QLD 4072, Australia, †Department of Cell and Molecular Biology, Faculty of Biotechnology and Biomolecular Sciences, Universiti Putra Malaysia, 43400, Serdang, Selangor, Malaysia, ‡Australian Centre for Plant Functional Genomics, School of Agriculture and Food Science, The University of Queensland, St Lucia, Brisbane QLD 4072, Australia, §Centre of Excellence in Genomics (CEG), International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru 502324, Telangana, India, and **School of Plant Biology, University of Western Australia, Crawley, WA 6009, Australia

ABSTRACT Genetic structure can be altered by chemical mutagenesis, which is a common method applied in molecular biology and genetics. Second-generation sequencing provides a platform to reveal base alterations occurring in the whole genome due to mutagenesis. A model legume, *Lotus japonicus* ecotype *Miyakojima*, was chemically mutated with alkylating ethyl methanesulfonate (EMS) for the scanning of DNA lesions throughout the genome. Using second-generation sequencing, two individually mutated third-generation progeny (M3, named AM and AS) were sequenced and analyzed to identify single nucleotide polymorphisms and reveal the effects of EMS on nucleotide sequences in these mutant genomes. Single-nucleotide polymorphisms were found in every 208 kb (AS) and 202 kb (AM) with a bias mutation of G/C-to-A/T changes at low percentage. Most mutations were intergenic. The mutation spectrum of the genomes was comparable in their individual chromosomes; however, each mutated genome has unique alterations, which are useful to identify causal mutations for their phenotypic changes. The data obtained demonstrate that whole genomic sequencing is applicable as a high-throughput tool to investigate genomic changes due to mutagenesis. The identification of these single-point mutations will facilitate the identification of phenotypically causative mutations in EMS-mutated germplasm.

KEYWORDS
abscisic acid
Lotus japonicus
mutagenesis
second-generation sequencing
SNP

Mutagenesis provides a powerful technique to improve plant breeding and assist functional and genomic analyses of crop plants. This technique was first introduced with the use of x-ray and radium radiations followed by fast neutron and gamma radiation (as reviewed in Sikora *et al.* 2011). Because such application of physical mutagens required specialized equipment, chemical mutagens were introduced later. Chemical mutagens are used widely because they are easier to handle and increase mutation frequency (Loveless 1958; Sikora *et al.* 2011; Serrat *et al.* 2014). Various chemical mutagens have been prepared, such as sodium azide, ethyl methanesulfonate (EMS), and *N*-ethyl-*N*-nitrosourea, which produce different side effects on the genetic structure of treated populations. These chemicals can cause point mutations, insertions, and/or deletions in the genomic strands, leading to phenotypic changes, which could be desirable traits for important crops (Olsen *et al.* 1993; Greene *et al.* 2003; Flibotte *et al.* 2010).

EMS, an alkylating agent, commonly is used as a chemical mutagen for DNA lesions. Unlike *N*-ethyl-*N*-nitrosourea, EMS induces a biased spectrum of G/C-to-A/T transitions. These transitions most likely occur due to the alkylation at the O^6 or N^7 position of guanine, which leads to the replacement of cytosine with thymine base pairing (Lawley and Martin 1975; Sega 1984; Haughn and Somerville 1987; Sikora *et al.* 2011). Known as EMS canonical base substitutions, the high frequency of G/C-to-A/T changes has been observed upon EMS

doi: 10.1534/g3.114.014571
Manuscript received October 30, 2014; accepted for publication February 2, 2015; published Early Online February 6, 2015.

[1]Corresponding author: Centre for Integrative Legume Research, School of Agriculture and Food Science, The University of Queensland, St Lucia, Brisbane QLD 4072, Australia. E-mail: p.gresshoff@uq.edu.au

exposure in different organisms, including *Arabidopsis thaliana* (Greene *et al.* 2003; Till *et al.* 2011), *Oryza sativa* (Till *et al.* 2011), *L. japonicus* (Perry *et al.* 2009), *Caenorhabditis elegans* (Flibotte *et al.* 2010; Thompson *et al.* 2013), *Solanum lycopersicum* (Minoia *et al.* 2010), and *Saccharomyces cerevisiae* (Shiwa *et al.* 2012) at different rates. EMS also tends to produce random point mutations and induces a low level of chromosomal breaks and lethal effects (Greene *et al.* 2003). These effects provide a competent survival rate and allow subsequent analyses to be performed for both forward and reverse genetics.

The whole-genome sequence of *Lotus japonicus* ecotype *Miyakojima* (MG-20) is available, covering a total length of 315,073,275 bp (67% of the 472-Mb genome). A total of 91.3% of gene space is located in the genome sequence (Sato *et al.* 2008). The genome sequences of the chloroplast (150,519 bp) and mitochondrion (380,861 bp) also have been assembled, by Kato *et al.* (2000) and Kazakoff *et al.* (2012), respectively. Numerous bioinformatics materials on legume and nonlegume plants are also publicly available from various resources (Sato and Tabata 2005; Goodstein *et al.* 2012; Li *et al.* 2012). With the current high-throughput tools, genome sequencing can be performed at an affordable cost (Thudi *et al.* 2012). Many programs also have been developed for analyzing sequencing data *in silico*, offering good strategies to employ our study.

In this study, we applied Illumina second-generation sequencing (2GS) to discover EMS effects and the mutation spectrum in the genome of model legume, *L. japonicus* ecotype MG-20. We randomly selected two mutagenized plant genotypes from an M3 population as our subjects that were deep-sequenced. Wild-type (WT) MG-20 also was resequenced and mapped to the reference genome from Sato *et al.* (2008) as a comparison to subtract natural variations and false positives. We aimed to scan the effects of EMS throughout the whole genome, regardless of the phenotypic characteristic, which resulted from the mutations in specific regions. We identified single-nucleotide polymorphisms (SNPs) and compared the base alterations that occurred between both the genomes. The data demonstrate how 2GS works as a high-throughput platform for rapidly identifying DNA changes in each EMS-induced genome. As an advantage over sequencing pooled mutants, scanning individual mutagenized genomes allows rapid analysis of the mutation spectrum and gives the actual picture of "corrupted" genetic structure. The output of this study also will provide information in the identification of genes mutated due to EMS mutagenesis and demonstrate the distribution of mutation is comparable in different germplasm of MG-20.

MATERIALS AND METHODS

Plant materials

A total of 4920 seeds of MG-20 were treated with 0.5% (v/v) EMS and grown as described in Biswas *et al.* (2009). One hundred MG-20 seeds were soaked in sterile water as a germination control. After 3 wk, the phenotypes of the plants that survived (67.3%) were observed to examine physical effects on the plant growth. Physiological tests also were performed as reported by Biswas *et al.* (2009) on abscisic acid (ABA) insensitivity and by Chan *et al.* (2013) on ethylene insensitivity. In this study, two mutated germplasms (M3) were isolated from the ABA assay (called AM and AS) for sequencing to screen EMS effects on their genome sequences and compare them with the resequenced WT MG-20 genome. In short, AM is a homozygous dominant, ABA-insensitive mutant (confirmed by stability in segregating families), whereas AS is the WT ABA phenotype segregant that was generated from a self-regeneration of a heterozygous ABA-insensitive mutant. Thus, AM and AS should carry the same spectrum of SNPs due to EMS mutagenesis.

DNA extraction

Genomic DNA was extracted from each individual plant after 1 mo of growth. The cetyl trimethylammonium bromide (CTAB) extraction method was adapted from Stewart and Via (1993) with minor modification. Plant tissues (about 1 g) were ground to powder in liquid nitrogen before adding 1 mL of CTAB extraction buffer. The mixture was incubated at 65° for 30−60 min. Five-hundred microliters of the mixture was transferred into a new 2-mL tube and 500 μL of CTAB buffer was added to each tube. Both tubes were incubated as before. Chloroform purification was performed followed by isopropanol precipitation and washing with 70% ethanol. The nucleic acid containing pellet was air dried and dissolved in 100 μL of sterile water and subjected to RNase treatment (Sambrook and Russel 2001). Extracted genomic DNA was assessed using a spectrophotometer and agarose gel electrophoresis.

Sequencing and SNP identification

Whole-genome paired-end, 100-bp, short-sequence reads (>10× coverage) for AM, AS, and WT MG-20 were generated using the Illumina Genome Analyzer IIx according to the manufacturer's instructions. These three datasets were then mapped to the MG-20 reference genome (www.kazusa.or.jp/lotus/) using SOAP2 v2.21 with the option –r0 to retain only uniquely mapping read pairs (Li *et al.* 2009). SNPs were called using SGSautoSNP 2.001 (Lorenc *et al.* 2012) wherein AM, AS, and WT were referred as different cultivars. To avoid false-positive output, only homozygous SNPs were selected for further analysis. The three datasets of paired reads have been deposited into the National Center for Biotechnology Information Short Read Archive database under accession: SRX719550 (AM), SRX729747 (AS), and SRX131060 (WT). All custom scripts are available on request by E-mail at dave.edwards@uq.edu.au.

SNP analysis

The resequenced WT genome was used for comparison to identify variants or base changes in the AS and AM genome sequences. The frequency of transitions and transversions generated in both mutated germplasms was also calculated. SNPs were categorized using SnpEff 3.0j (Cingolani *et al.* 2012) according to their effect on *L. japonicus* MG-20−annotated genes (Sato *et al.* 2008).

Selection of genes involved in ABA perception and signaling pathways

A number of ABA candidate genes also were selected to test whether a SNP is located in their sequences. As a preliminary study, a total of 32 genes reported to be involved in ABA signaling were selected as candidate genes. These genes commonly are reported in the ABA-gene interaction in ABA perception and signaling pathways (Wasilewska *et al.* 2008; Cutler *et al.* 2010; Kim *et al.* 2010).

RESULTS

Phenotypic effects

All control seeds successfully germinated and were well-developed. EMS-treated seeds had a germination rate of 67.3% (3313/4920). Among well-developed EMS-treated plants, 76 M1 plants showed abnormal phenotypes after 3 wk of growth. Almost 1% of these plants (30/3313) showed an albino phenotype, in which yellow sectors were observed. Some plants also had pale green patches (0.6%, 23/3313), early branching (0.2%, 7/3313) or a looped base (0.39%, 13/3313). Two other plants showed either unusual leaf shape or early flowering. Vivipary also was detected on a pod among this population. In this

study, the impaired phenotypes of *L. japonicus* indicated the successful treatment of mutagenesis using 0.5% EMS, and therefore, the population could be utilized for subsequent analyses. We chose mutants at the third generation (M3) as subjects to ensure the mutation is stable and fixed (Serrat *et al.* 2014).

Sequencing and read mapping output

A total of 32,965,291, 34,020,296, and 25,737,274 paired reads were generated from the WT, AS, and AM genomes, respectively (Table 1). At least 23% of reads from the sequenced genomes mapped to the reference. The percentage of mapped read pairs was 28.17%, 30.28%, and 23.59% from WT, AS, and AM, respectively, which resulted in more than 19× genome coverage. As a result, the genome coverage was sufficient to be applied for SNP calling between mutants and WT reads.

Frequency of mutation

After read mapping, homozygous SNPs were predicted, to identify SNPs that are unique in the mutant genome as opposed to WT. As a result, the frequency of mutation could be observed in assembled sequences of all chromosomes (Table 2). Chromosomes 1 and 6 have the longest and shortest length of assembled sequences, respectively. Meanwhile, the unmapped contigs cover approximately 32.9 Mb of total assembled length. Our SNP calling showed that Chromosome 1 has the greatest number of SNPs with one mutation per 170 kb and 165 kb in the AS and AM genomes, respectively. Mutation frequency was the lowest in Chromosome 6, which had less than 6% of total SNP in both genomes. The change rates were one mutation per 222 kb for AS and 217 kb for AM. In total, mutation frequency in AM and AS was nearly identical 1490 SNPs and 1447 respectively. This frequency is reflected in the change rate of mutation in the AM genome (one homozygous mutation in every 202 kb) and AS genome (one homozygous mutation in every 208 kb).

The frequency of transitions and transversions generated was analyzed (Table 3). In our mutants, EMS has generated 64.7% (AS) and 62.3% (AM) transitions as a percent of total mutation. Both mutants showed a bias to G/C-to-A/T transitions, which were the most frequent mutations that occurred (45.0% in AS and 34.9% in AM). Transversion mutation was also detected, but at lower percentages as listed in Table 3. The lowest percentage of mutation was C/G-to-G/C changes in AS and A/T-to- T/A changes in AM.

Distribution of transition and transversion mutations

Based on the mutation frequency, how each mutation type was located in individual chromosomes can be observed by calculating the percentage of mutation type in relation to the SNP total in each chromosome (Figure 1). Both genomes comprised of C/T or G/A transitions as the most frequent mutation type in their chromosomes or unmapped regions. In the AS genome, G/A transitions were the highest in Chromosomes 1, 2, 5, and 6. Chromosomes 3 and 4 and unmapped regions had C/T transitions as the highest percentage of mutation type. The lowest percentage of mutation type was C/G transversions which were present the least in all individual chromosomes and unmapped regions of the AS mutant. The distribution of transitions and transversions in the AM genome was relatively similar to the AS genome. The greatest percentage of mutation type in each chromosome of the AM genome was C/T transitions except Chromosome 3. Unmapped regions and Chromosome 3 of AM have G/A transitions as the greatest percentage of mutation type. Meanwhile, the lowest percentage of mutation type was detected to be either A/T or C/G transversions in the AM genome.

Distribution of SNPs across mutagenized genomes

We determined how many SNPs occurred every 1000 kb to show the distribution of mutations across our mutagenized genomes (Figure 2). Regardless of mutation types, the distribution of SNPs is unique between AS and AM when comparing the same chromosome. SNPs were randomly located along the genomes with no specific chromosome position being particularly abundant or lacking in SNPs for both genomes. However, Chromosomes 1 and 2 were highly "corrupted" in their arms. Meanwhile, some chromosomes (Chromosomes 3, 4, and 5 for AS; Chromosome 3 for AM) have a high peak of SNPs toward their center. The density of mutations in every 1000 kb was relatively apparent in Chromosomes 1 and 2 of AS and AM. Meanwhile, Chromosome 6 had less dense mutations for both genomes.

Coding and noncoding effects

Here, we used SnpEff to predict the effect of mutations on coding regions of annotated genes of *L. japonicus* for both mutants, as shown in Figure 3. Our results showed that the greatest number of SNPs (34%) was predicted to be located in intergenic regions, followed by downstream and upstream regions of predicted genes (27% each). Only 7% and 3% of SNPs were located in exon and intron sequences, respectively. Intragenic regions also were predicted to have low EMS effect. The mutation percentage was very low at splice site donor and acceptors. In our mutagenized genomes, EMS effects did not contribute to a high fraction of nonsynonymous and synonymous changes (5% and 2%, respectively).

Candidate genes

As preliminary research, we selected 32 notorious ABA genes (Table 4) to determine whether any SNPs were located in their sequences. Orthologous genes of *L. japonicus* were identified based on *A. thaliana* and *Glycine max* sequences. Six were found to have orthologous sequences, which were positioned in unmapped regions of the MG-20 genome. The rest of the orthologs were found in the assembled chromosomes. SNPs were identified in three candidate loci, which were ABI2, ABI3, and ABI4. ABI2 and ABI4 were mutated in both mutants, whereas SNP was only found in ABI3 of the AS mutant.

DISCUSSION

The effect of EMS mutagenesis initially can be observed through phenotypic changes of mutagenized plants, and numerous reports

Table 1 Outputs generated from Illumina sequencing to read mapping

Genome	Paired Raw Reads	Read Pairs Mapped	% of Read Pairs Mapped	Genome Coverage[a]
MG-20 WT	32,965,291	9,285,440	28.17	29.88X
AS mutant	34,020,296	10,299,840	30.28	33.15X
AM mutant	25,737,274	6,071,230	23.59	19.54X

WT, wild type.
[a] Based on mapped reads.

Table 2 Frequency of SNPs in individual chromosomes and unmapped regions of AS and AM mutants

		AS Genome		AM Genome	
Chromosome	Length, bp	Change (SNPs)	Change Rate	Change (SNPs)	Change Rate
1	66,776,104	391	170,783	404	165,287
2	44,510,304	258	172,521	250	178,041
3	48,258,781	208	232,013	204	236,563
4	43,347,107	176	246,290	195	222,293
5	37,320,184	190	196,422	200	186,601
6	28,216,978	76	371,276	86	328,104
Unmapped	32,912,371	148	222,381	151	217,963
Total	301,341,829	1447	208,253	1490	202,243

Total of assembled length (bp), base changes, and change rate of individual chromosomes and unmapped regions are listed in the table. SNPs, single-nucleotide polymorphisms.

have been published on the effects. The abnormal phenotypic variance due to EMS exposure also has been reported previously in *L. japonicus* (Perry *et al.* 2003) and other plants such as *Glycine max* (soybean; Carroll *et al.* 1986), *Hordeum vulgare* L. (barley; Caldwell *et al.* 2004), *Sorghum bicolor* (L.) Moench Xin *et al.* 2008), *Solanum lycopersicum* (tomato; Minoia *et al.* 2010), and *Capsicum annuum* L. (chilli pepper; Sri Devi and Mullainathan 2012). The percentage of the abnormal phenotypes varied in different plants depending on the concentration of EMS used (Xin *et al.* 2008). Phenotypic changes are not restricted to plant growth but also affect their physiological characteristics. For examples, reduced sensitivity to ABA was reported by Biswas *et al.* (2009) in *L. japonicus* mutagenized by EMS. Response to high temperature also was affected in EMS-induced *Arabidopsis* mutants impaired in ABA and salicylic acid syntheses (Zhu *et al.* 2012).

Sequencing data often are used to mine SNPs or polymorphisms among different cultivars for identification of traits and allelic variations in crops (Varshney *et al.* 2012; Edwards *et al.* 2013; Zander *et al.* 2013). The high-throughput technology also is applied in discovering causative mutations by pooling backcrossed segregant populations to increase SNP frequency and reveal mutated regions (Ashelford *et al.* 2011; Mokry *et al.* 2011; Hartwig *et al.* 2012). Here, we individually sequenced two selected mutant genomes and used 2GS data to compare mutation distribution between these mutants.

To our knowledge, this is the first report on the application of SGSautoSNP for discovery of SNPs induced by EMS mutagenesis. This tool uses assembled reads to identify homozygous SNPs without using the reference genome, which is only used for mapping reads. SGSAutoSNP was used to identify only homozygous SNPs because the identification of heterozygote SNPs leads to a high number of false positives, which would lead to an overrepresentation of the mutation frequency in the genome (Lorenc *et al.* 2012). We also managed to rapidly identify the mutation spectrum occurring in the mutant genomes using the output data.

Our read mapping has been successfully performed before SNP identification in each genome. However, a low percent of reads was mapped to our reference. A high percentage of paired reads was reported to map to the reference in other species such as *Caenorhabditis elegans* (average 92%; Zuryn *et al.* 2010) and *A. thaliana* (average 73%; Austin *et al.* 2011). In our case, only an average of 27% paired reads mapped. Similar results were observed in other species using the same SNP prediction software [for example in canola (Dalton-Morgan *et al.* 2014) and wheat (Lai *et al.* 2015)]. The number of reads mapped relies on the parameters set up during the mapping procedure. We only selected paired reads that mapped and ignored reads that aligned as single reads and only used reads that mapped to a single unique location to generate more accurate read mapping (Li *et al.* 2009; Lorenc *et al.* 2012). The assembled length of pseudomolecules in our MG-20 reference covers 67% of estimated genome size (Sato *et al.* 2008), which also influences read mapping output. In addition, reads that aligned to multiple positions also were removed to increase SNP calling accuracy and avoid false positives (Lorenc *et al.* 2012; Shiwa *et al.* 2012). These factors have reduced the number of reads mapped in this study.

Chromosome 3 is the largest among the six chromosomes of the MG-20 genome. However, the assembled length of Chromosome 1 is the longest, and presumably more complete, based on the genome assembly (Sato *et al.* 2008), allowing for better read mapping and a greater rate of SNP identification. This reflects the greatest number of SNPs predicted in Chromosome 1. Chromosome 6 has the lowest number of identified SNPs, because it is the shortest chromosome in assembled length (Sato *et al.* 2008). In addition, Chromosome 6 might contain more repetitive sequences and therefore less unique reads mapped, leading to a lower rate of SNPs being identified. Furthermore, the different rates of base changes between chromosomes may demonstrate the different capacity levels of each chromosome in tolerating mutation impact. There is insufficient evidence to reach clear conclusions; however, the presence of selection pressure, such as clustering of housekeeping genes on chromosomes, could contribute to inability to tolerate mutations. Additionally, not all SNPs will be identified across the genome due to repetitive regions and the stringency of read mapping, reflecting SNP/mutation density across chromosomes.

A different number of SNPs between AS and AM was expected, because it is impossible to obtain an exact total of SNPs in different mutated genomes. The mutation rate of our mutants was 1/208 kb and 1/202 kb, which were greater than previously reported in *L. japonicus* (1/502 kb; Perry *et al.* 2009). Different rates of mutation also were reported in various plants mutagenized by EMS, as summarized by Sikora *et al.* (2011). Although only homozygous SNPs

Table 3 Spectrum of base mutation found in the AS and AM genomes

	Changes, %	
Mutation	AS	AM
Transition		
G/C-to-A/T	45.0	34.9
A/T-to-G/C	19.7	27.4
Transversion		
A/C-to-C/A	12.7	15.3
G/T-to-T/G	12.6	13.3
A/T-to- T/A	6.9	4.5
C/G-to-G/C	3.1	4.6

A high frequency of transition mutation was observed as expected

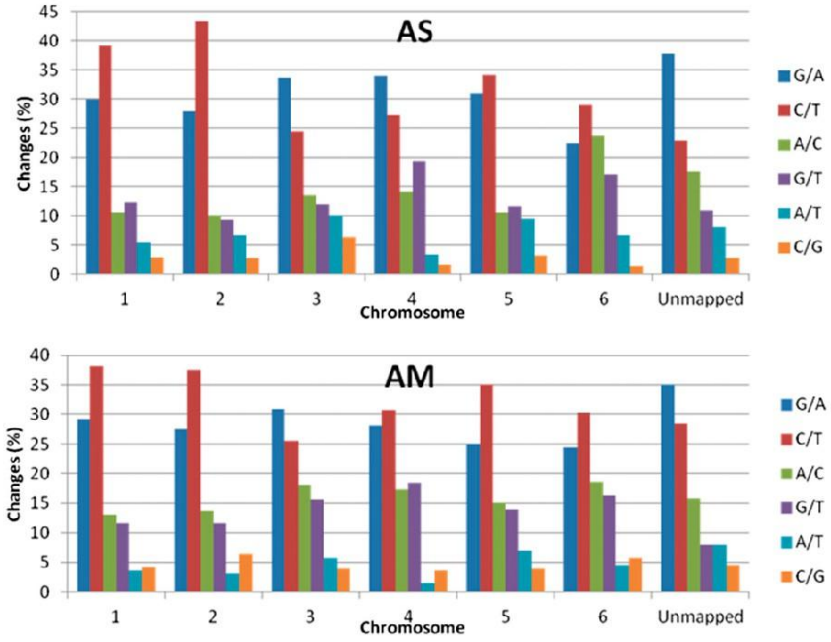

Figure 1 Relative percentage of different mutation types. Our mutated genomes had varied percentages of base changes in each chromosome and unmapped regions. A high percentage of G/A and C/T changes was observed in both genomes.

were taken into account in this study, the frequency of mutation was quite high. If heterozygous mutations were called, the frequency will be greater. Here, the frequency of mutation demonstrates the effectiveness of EMS to produce high mutagenesis as reported by many researchers (Dube *et al.* 2011; Shiwa *et al.* 2012; Serrat *et al.* 2014).

The mutation spectrums in our germplasms were not consistent with a previous report on TILLING work of mutagenized *L. japonicus Gifu* (Perry *et al.* 2009) wherein 97.6% of base changes were G/C-to-A/T mutation upon EMS mutagenesis. Similar to TILLed *Arabidopsis*, a high frequency of G/C-to-A/T transition (99%) also was identified (Greene *et al.* 2003). A lower rate of G/C-to-A/T base changes was observed in other plants, including barley (70%; Caldwell *et al.* 2004), rice (70%; Till *et al.* 2007), and tomato (60%; Minoia *et al.* 2010), showing a possibility of the presence of various base changes upon EMS exposure. However, transition mutations could not be denied as the most common base mutation in the EMS sphere as reported previously (Lawley and Martin

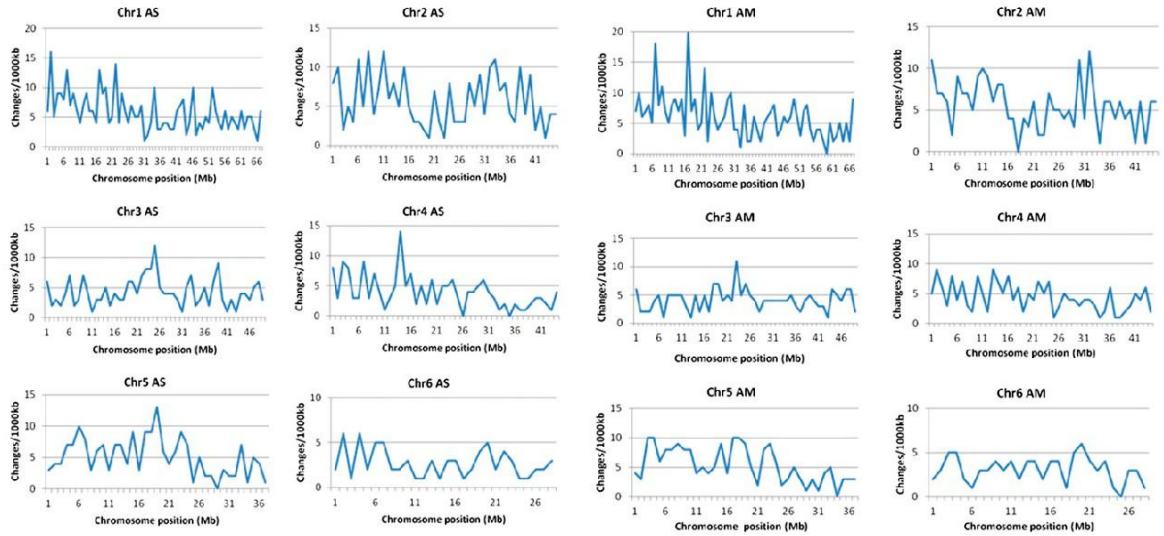

Figure 2 Distribution of mutation across individual chromosomes in both AS (left) and AM (right) genomes. Mutations were plotted in every 1000 kb of genomic sequences. Chromosome (Chr) number is shown above its designated graph. Chromosome position was plotted based on the assembled length of *L. japonicus* genome.

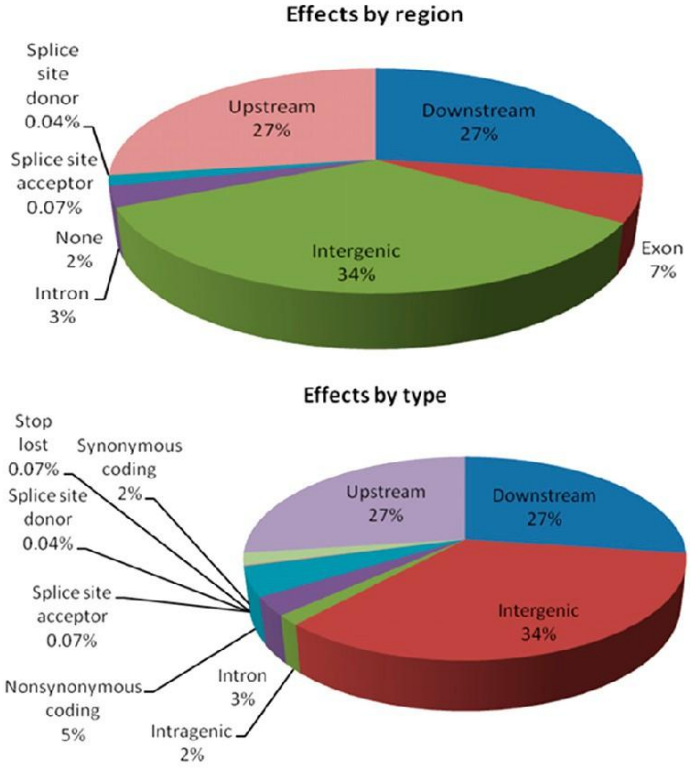

Figure 3 Mutation effects on codon sequences by type and region in our mutant genomes. Single-nucleotide polymorphisms were observed highly located in intergenic regions, upstream, and downstream parts of the annotated genes. Only a small percentage of nonsynonymous changes was predicted.

1975; Sikora *et al.* 2011; Shiwa *et al.* 2012). In the TLLING analysis, a number of interesting genes with known sequences were selected for the identification of base changes in pooled mutated genomes (McCallum *et al.* 2000). In this study, we adopted a high-throughput technology (2GS) to scan homozygous SNPs that are present throughout the individual mutated genomes compared with our resequenced WT MG-20. This provided a wider range of EMS effects in a genome. Our approach also neglected natural variation between the reference and the resequenced MG-20 genomes. Our reference was used merely for read mapping and has not been counted during SNP calling (Lorenc *et al.* 2012).

Although the percentage of transversion mutations was relatively low in both genomes, they should not be omitted in a mutation study because they potentially are causal mutations for phenotypes of interest. Taking into account only homozygous mutations, our data also show that the presence of specific base changes is comparatively distributed among individual chromosomes and not abundantly located in a specific chromosome for both alleles. Comparing these two genomes, we found that transition and transversion mutations were present nearly consistently in the same chromosomes. These results reveal that the frequency of EMS-induced transition and transversion mutations is comparable between different individual mutants that were derived from the same mutagenized population.

As mentioned previously, the total assembled length affects the SNP total identified in each chromosome and consequently, their distributions. The longer the assembled length, the denser the mutation distribution was found to be along the chromosome. The mutation distribution also was affected by the occurrence of assembled sequences at specific regions (www.kazusa.or.jp/lotus/). Chromosome 6 is the best example, in which low SNP peaks were detected and the mutations distributed broadly. Read mapping is very difficult in centromeric regions, and therefore the density in these regions has not been investigated in this study. Distribution of mutations was scattered due to the effect of EMS as a mutagen that causes a random mutation (Lawley and Martin 1975; Greene *et al.* 2003; Sikora *et al.* 2011; Tagu *et al.* 2014). A random distribution also was reported by Thompson *et al.* (2013) in the genome of single *C. elegans* from different strains, which were mutagenized by EMS. They propagated mutagenized worms through 10 generations to obtain stable mutants. However, this is time-consuming for legumes, which have a longer generation period. Nevertheless, our 2GS data of third-generation mutants also could provide the literal effect of EMS mutagenesis throughout the individual genomes.

EMS induces base changes or nucleotide substitution, which consequently alter codon sequences, leading to either nonsynonymous or synonymous effects. In genetics studies, nonsynonymous change is a favorable mutation effect because it gives a clue on which gene may be associated with a specific phenotype (Ng and Henikoff 2006). Our results showed that EMS has arbitrary and broad effects of mutation on codon sequences. Mutated coding regions could represent nonsynonymous changes, which are useful to identify gene of interest for desirable phenotypes.

To extend the implication of our data, preliminary work has been employed to specifically detect the presence of mutations in our genes of interest. A total of 32 genes reported to be involved in ABA signaling were selected as candidate genes. These genes are commonly reported in the ABA−gene interaction in ABA perception and signaling pathways (Wasilewska *et al.* 2008; Cutler *et al.* 2010; Kim *et al.* 2010). We chose to use ABA candidate genes because our mutants

Table 4 List of ABA candidate genes and their loci in the genome of *Arabidopsis*, soybean, and *Lotus*

No.	Candidate Gene	Molecular Function	Locus		
			Arabidopsis thaliana	*Glycine max*	*Lotus japonicus*
1	*ABA INSENSITIVE 2* (*ABI2*)	Protein phosphatase 2C	AT5G57050	Gm17g33410.1	chr1.CM0133.740.r2.m
2	*ABI3*	Transcription factor	AT3G24650	Gm08g47240.1	chr1.CM0147.920.r2.d
3	*ABI4*	Transcription factor	AT2G40220	Gm02g31350.1	chr1.CM0318.160.r2.d
4	*ABI5*	bZIP transcription factor	AT2G36270	Gm19g37910.1	chr1.CM0010.100.r2.d
5	*ABI8*	Glycosyl transferases	AT3G08550	Gm04g26230.1	chr3.CM2163.130.r2.m
6	*AGB1*	Heterotrimeric G-protein complex	AT4G34460.1	Gm11g12600.1	chr1.CM0113.1970.r2.d
7	*AHA1/ OPEN STOMATA 2* (*OST2*)	ATPase	At2G18960.1	Gm13g00840.1	chr4.CM0244.50.r2.m
8	*ABA Responsive Element-binding Factor2* (*ABF2*)/*AREB1*	bZIP transcription factor	AT1G45249.3	Gm06g04350.1	chr1.CM2113.380.r2.a
9	*ATPT2/ PHT1*:4	Phosphatase transporter	AT2G38940	Gm19g34710.1	chr1.CM0295.140.r2.m
10	*Ethylene Response DNA binding factor 3* (*EDF3*)	Transcription factor	AT3G25730	Gm10g34760.1	LjSGA_080421.2
11	*Ethylene Response Factor 7* (*ERF7*)	Transcription factor	At3G20310	Gm14g02360.1	LjSGA_013296.1.1
12	*Enhanced Response to ABA 1* (*ERA1*)	Farnesyltransferase activity	AT5G40280	Gm13g23780.1	chr2.CM0081.550.r2.d
13	FUS3	Transcription factor	AT3G26790.1	Gm19g27340.1	chr1.CM0104.400.r2.a
14	*Gated Outwardly Rectifying K+ Channel* (*GORK*)	Outward rectifier K channel activity	AT5G37500.1	Gm02g41040.1	chr6.CM0508.670.r2.m
15	*Stelar K+ Outward Rectifier* (*SKOR*)	Outward rectifier K channel activity	AT3G02850	Gm14g39330.1	Same as GORK
16	*GPA1*	GTP binding protein	AT2G26300.1	Gm06g05960.1	chr5.CM0034.250.r2.m
17	*Glutathione peroxidase 3* (*GPX3*)	Redox transducer & scavenger	AT2G43350.1	Gm11g02630.1	chr4.CM0004.300.r2.m
18	*GPCR-TYPE G protein 1* (*GTG1*)	ABA receptor	AT1G64990.1	Gm12g0174	chr3.CM0127.40.r2.m
19	*High Leaf temperature 1* (*HT1*)	Serine/threonine protein kinase	AT1G62400	Gm07g39460.1	chr4.CM0288.800.r2.m
20	*Keep On Going* (*KEG*)	Ring E3 ligase	AT5G13530	Gm11g25680.1	LjSGA_007856.1
21	*Potassium Channel In A. thaliana 1* (*KAT1*)	Cyclic-nucleotide binding	AT5G46240	Gm08g24960.1	chr6.CM157.280.r2.a
22	*OST1*	Serine/threonine protein kinase	AT4G33950.1	Gm02g15330.1	LjSGA_038133.1
23	*PLDα1*	Phospholipase D/ transphosphatidylase	AT3G15730.1	Gm07g03490.1	chr3.CM0142.570.r2.d
24	*Pyrabactin Resistance1-like 2* (*PYL2*)/*Regulatory Component of ABA Receptor 14* (*RCAR14*)	Polyketide cyclase/dehydrase	AT2G26040	Gm04g05380.1	LjSGA_056222.1
25	*PYL3/RCAR13*	Polyketide cyclase/dehydrase	Similar to PYL2	Similar to PYL2	Similar to PYL2
26	*PYL4/RCAR10*	Polyketide cyclase/dehydrase	AT2G38310	Gm18g37410.1	chr3.CM0116.270.r2.m
27	*PYL5/RCAR8*	Polyketide cyclase/dehydrase	AT5G05440	Gm02g42990.1	LjSGA_020312.1
28	*Pyrabactin Resistance1* (*PYR1*)/ *Regulatory Component of ABA Receptor 11* (*RCAR11*)	*Streptomyces* cyclase/ dehydrase	AT4G17870.1	Gm08g36770.1	chr2.CM0177.730.r2.m
29	*Regulator of G protein Signaling 1* (*RGS1*)	G-protein coupled receptor	AT3G26090	Gm11g37540.1	chr6.LjT45M05.110.r2.d
30	*TPK10/CIPK15*	CBL-interacting serine/ threonine protein kinase	AT5G01810.1	Gm18g44450.1	chr3.LjT45I18.90.r2.d
31	*Slow Anion Channel Associated 1* (*SLAC1*)	C4-dicarboxylate transporter/ malic acid transport	AT1G12480	Gm09g23220.1	LjSGA_063103.1
32	*MYB44*	Transcription factor	None	Gm04g04490.1	chr5.CM0096.100r2.m

ABA, abscisic acid.

were isolated from a mutagenized population effected in response to ABA. We did identify base changes in several candidate genes, indicating a low rate of EMS effects on these sequences. In addition, identified SNPs were not only in the ABA-insensitive AM mutant, indicating they were background mutations. On the other hand, comparing mutations between AM and AS mutants would provide clues on the causal mutation of mutant phenotypes. However, other candidate genes also can be considered because ABA is involved in a wide range of plant growth and responses to environmental stresses. Its effects are varied in regulating the events of plant physiology, which require gene interaction and/or cross talks with other hormones (as reviewed in Kermode 2005; Fujita *et al.* 2006; Rock *et al.* 2010). Because this research worked on an individual genome of mutants, another strategy has been developed to sequence pooled DNA from different mutants to remove the effect of background mutation and increase the likelihood of identifying the causative gene.

In conclusion, this scanning revealed a detailed effect of EMS mutation on the whole genome of an individual mutant. Our results presented an overview of point mutations that occurred in the genome of mutants, which were usually pooled to identify SNPs. Here, EMS has produced a number of abnormal plants in our mutant population of *L. japonicus*. Our 2GS data also revealed how EMS efficiently mutates genomic sequences in an individual mutagenized plant. As expected, EMS created a random spectrum of mutation across the whole genome of our mutants and biased to G/C-to-A/T changes. However, other transition and transversion mutations also were identified with quite apparent fractions. Calling only homozygous SNPs has put a high confidence in identified base changes occurred. Mutation distribution apparently is distinct between mutated genomes, which derived from the same mutagenised population. Effects of SNPs on coding and noncoding regions could be manipulated to identify a causal mutation of a phenotype of interest. Our next question will be which gene is responsible for our abnormal phenotype of interest. This 2GS data will be further analyzed to discover the causative mutated gene.

ACKNOWLEDGMENTS

We thank the Centre for Integrative Legume Research, the University of Queensland for giving the opportunity to run this work. Special thanks to Dr Stephen Kazakoff, Queensland Centre for Medical Genomics, the University of Queensland for the technical support. A scholarship from the Ministry of Higher Education, Malaysia, in financially supporting our student also is acknowledged.

LITERATURE CITED

Ashelford, K., M. E. Eriksson, C. M. Allen, R. D'Amore, M. Johansson *et al.*, 2011 Full genome re-sequencing reveals a novel circadian clock mutation in *Arabidopsis*. Genome Biol. 12: R28.

Austin, R. S., D. Vidaurre, G. Stamatiou, R. Breit, N. J. Provart *et al.*, 2011 Next-generation mapping of Arabidopsis. Plant J. 67: 715–725.

Biswas, B., P. K. Chan, and P. M. Gresshoff, 2009 A novel ABA insensitive mutant of *Lotus japonicus* with a wilty phenotype displays unaltered nodulation regulation. Mol. Plant 2: 487–499.

Caldwell, D. G., N. McCallum, P. Shaw, G. J. Muehlbauer, D. F. Marshall *et al.*, 2004 A structured mutant population for forward and reverse genetics in barley (*Hordeum vulgare* L.). Plant J. 40: 143–150.

Carroll, B. J., D. L. McNeil, and P. M. Gresshoff, 1986 Mutagenesis of soybean (*Glycine max* (L.) Merr) and the isolation of non-nodulating mutants. Plant Sci. 47: 109–114.

Chan, P. K., B. Biswas, and P. M. Gresshoff, 2013 Classical ethylene insensitive mutants of the *Arabidopsis EIN2* orthologue lack the expected 'hypernodulation' response in *Lotus japonicus*. J. Integr. Plant Biol. 55: 395–408.

Cingolani, P., A. Platts, L. L. Wang, M. Coon, T. Nguyen *et al.*, 2012 A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w^{1118}; *iso*-2; *iso*-3. Landes Bioscience 6: 1–13.

Cutler, S. R., P. L. Rodriguez, R. R. Finkelstein, and S. R. Abrams, 2010 Abscisic acid: emergence of a core signalling network. Annu. Rev. Plant Biol. 61: 651–679.

Dalton-Morgan, J., A. Hayward, S. Alamery, R. Tollenaere, A. S. Mason *et al.*, 2014 A high-throughput SNP array in the amphidiploid species *Brassica napus* shows diversity in resistance genes. Funct. Integr. Genomics 14: 643–655.

Dube, K. G., A. S. Bajaj, and A. M. Gawande, 2011 Mutagenic efficiency and effectiveness of gamma rays and EMS in *Cyamopsis tetragonoloba* (L.) var. *Sharada*. Asiatic J. Biotechnol. Resour. 2: 436–440.

Edwards, D., J. Batley, and R. J. Snowdon, 2013 Accessing complex crop genomes with next-generation sequencing. Theor. Appl. Genet. 126: 1–11.

Flibotte, S., M. L. Edgley, I. Chaudhry, J. Taylor, S. E. Neil *et al.*, 2010 Whole-genome sequencing profiling of mutagenesis in *Caenorhabditis elegans*. Genetics 185: 431–441.

Fujita, M., Y. Fujita, Y. Noutoshi, F. Takahashi, Y. Narusaka *et al.*, 2006 Crosstalk between abiotic and biotic stress responses: a current view from the points of convergence in the stress signaling networks. Curr. Opin. Plant Biol. 9: 436–442.

Goodstein, D. M., S. Shu, R. Howson, R. Neupane, R. D. Hayes *et al.*, 2012 Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 40: D1178–D1186.

Greene, E. A., C. A. Codomo, N. E. Taylor, J. G. Henikoff, B. J. Till *et al.*, 2003 Spectrum of chemically induced mutations from a large-scale reverse genetic screen in Arabidopsis. Genetics 164: 731–740.

Hartwig, B., G. V. James, K. Konrad, K. Schneeberger, and F. Turck, 2012 Fast isogenic mapping-by-sequencing of ethyl methanesulfonate-induced mutant bulks. Plant Physiol. 160: 591–600.

Haughn, G., and C. R. Somerville, 1987 Selection for herbicide resistance at the whole-plant level, pp. 98–107 in *Biotechnology in Agricultural Chemistry, ACS Symposium Series*, edited by H. LeBron *et al.* American Chemical Society, Washington, DC.

Kato, T., T. Kaneko, S. Sato, Y. Nakamura, and S. Tabata, 2000 Complete structure of the chloroplast genome of a legume, *Lotus japonicus*. DNA Res. 7: 323–330.

Kazakoff, S. H., M. Imelfort, D. Edwards, J. Koehorst, B. Biswas *et al.*, 2012 Capturing the biofuel wellhead and powerhouse: the chloroplast and mitochondrial genomes of the leguminous feedstock tree *Pongamia pinnata*. PLoS ONE 7: e51687.

Kermode, A. R., 2005 Role of abscisic acid in seed dormancy. J. Plant Growth Regul. 24: 319–344.

Kim, T. H., M. Böhmer, H. Hu, N. Nishimura, and J. I. Schroeder, 2010 Guard cell signal transduction network: advances in understanding abscisic acid, CO_2 and Ca^{2+} signalling. Annu. Rev. Plant Biol. 61: 561–591.

Lai, K., M. T. Lorenc, C. L. Hong, P. J. Berkman, P. E. Bayer *et al.*, 2015 Identification and characterisation of more than 4 million intervarietal SNPs across the group 7 chromosomes of bread wheat. Plant Biotechnol. J. 13: 97–104.

Lawley, P. D., and C. N. Martin, 1975 Molecular mechanisms in alkylation mutagenesis. Induced reversion of bacteriophage T4rII AP72 by ethyl methanesulphonate in relation to extent and mode of ethylation of purines in bacteriophage deoxyribonucleic acid. Biochem. J. 145: 85–91.

Li, J., X. Dai, T. Liu, and P. X. Zhao, 2012 LegumeIP: an integrative database for comparative genomics and transcriptomics of model legumes. Nucleic Acids Res. 40: D1221–D1229.

Li, R., C. Yu, Y. Li, T. W. Lam, S. M. Yiu *et al.*, 2009 SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25: 1966–1967.

Lorenc, M. T., S. Hayashi, J. Stiller, H. Lee, S. Manoli *et al.*, 2012 Discovery of single nucleotide polymorphisms in complex genomes using SGSautoSNP. Biology 1: 370–382.

Loveless, A., 1958 Increased rate of plaque-type and host-range mutation following treatment of bacteriophage *in vitro* with ethyl methane sulphonate. Nature 181: 1212–1213.

McCallum, C. M., L. Comai, E. A. Greene, and S. Henikoff, 2000 Targeting induced local lesions in genomes (TILLING) for plant functional genomics. Plant Physiol. 123: 439–442.

Minoia, S., A. Petrozza, O. D'Onofrio, F. Piron, G. Mosca *et al.*, 2010 A new mutant genetic resource for tomato crop improvement by TILLING technology. BMC Res. Notes 3: 69.

Mokry, M., I. J. Nijman, A. van Dijken, R. Benjamins, R. Heidstra *et al.*, 2011 Identification of factors required for meristem function in *Arabidopsis* using a novel next generation sequencing fast forward genetics approach. BMC Genomics 12: 256.

Ng, P. C., and S. Henikoff, 2006 Predicting the effects of amino acid substitutions on protein function. Annu. Rev. Genomics Hum. Genet. 7: 61–80.

Olsen, O., X. Wang, and D. von Wettstein, 1993 Sodium azide mutagenesis: preferential generation of A·T → G·C transitions in the barley *An18* gene. Proc. Natl. Acad. Sci. USA 90: 8043–8047.

Perry, J. A., T. L. Wang, T. J. Welham, S. Gardner, J. M. Pike *et al.*, 2003 A TILLING reverse genetics tool and a web-accessible collection of mutants of the legume *Lotus japonicus*. Plant Physiol. 131: 866–871.

Perry, J., A. Brachmann, T. Welham, A. Binder, M. Charpentier *et al.*, 2009 TILLING in *Lotus japonicus* identified large allelic series for symbiosis genes and revealed a bias in functionally defective ethyl methanesulfonate alleles toward glycine replacements. Plant Physiol. 151: 1281–1291.

Rock, C. D., Y. Sakata, and R. S. Quatrano, 2010 Stress signaling I: The role of abscisic acid (ABA), pp. 33–73 in *Abiotic Stress Adaptation in Plants: Physiological, Molecular and Genomic Foundation*, edited by A. Pareek, S. K. Sopory, and H. J. Bohnert and Govindjee. Springer, Dortrecht.

Sambrook, J., and D. W. Russel, 2001 *Molecular Cloning: A Laboratory Manual*, Ed. 3. Cold Spring Harbor Laboratory Press, New York.

Sato, S., and S. Tabata, 2005 *Lotus japonicus* as a platform for legume research. Curr. Opin. Plant Biol. 9: 128–132.

Sato, S., Y. Nakamura, T. Kaneko, E. Asamizu, T. Kato *et al.*, 2008 Genome structure of the legume, *Lotus japonicus*. DNA Res. 15: 227–239.

Sega, G. A., 1984 A review of the genetic effects of ethyl methanesulfonate. Mutat. Res. 134: 113–142.

Serrat, X., R. Esteban, N. Guibourt, L. Moysset, S. Nogués *et al.*, 2014 EMS mutagenesis in mature seed-derived rice calli as a new method for rapidly obtaining TILLING mutant populations. Plant Methods 10: 5.

Shiwa, Y., S. Fukushima-Tanaka, K. Kasahara, T. Horiuchi, and H. Yoshikawa, 2012 Whole-genome profiling of a novel mutagenesis technique using proofreading-deficient DNA polymerase δ. Int. J. Evol. Biol. 2012: 860797.

Sikora, P., A. Chawade, M. Larsson, J. Olsson, and O. Olsson, 2011 Mutagenesis as a tool in plant genetics, functional genomics and breeding. Int. J. Plant Genom. 2011: 314829.

Sri Devi, A., and L. Mullainathan, 2012 The use of ethyl methanesulfonate to study the flower development in *Capsicum annuum* L. mutants. Bot. Res. Intl. 5: 4–9.

Stewart, C. N., and L. E. Via, 1993 A rapid CTAB DNA isolation technique useful for RAPD fingerprinting and other PCR applications. Biotechniques 14: 748–749.

Tagu, D., G.L. Trionnaire, S. Tanguy, J.P. Gauthier, and J.R. Huynh, 2014 EMS mutagenesis in the pea aphid *Acyrthosiphon pisum*. G3 (Bethesda) 4: 657–667.

Thompson, O., M. Edgley, P. Strasbourger, S. Flibotte, B. Ewing *et al.*, 2013 The million mutation project: a new approach to genetics in *Caenorhabditis elegans*. Genome Res. 23: 1749–1762.

Thudi, M., Y. Li, S. A. Jackson, G. D. May, and R. K. Varshney, 2012 Current state-of-art of sequencing technologies for plant genomics research. Brief Funct. Genomics 11: 3–11.

Till, B. J., J. Cooper, T. H. Ti, P. Colowit, E. A. Greene *et al.*, 2007 Discovery of chemically induced mutations in rice by TILLING. BMC Plant Biol. 7: 19.

Till, B. J., S. H. Reynolds, E. A. Greene, C. A. Codomo, L. C. Enns *et al.*, 2011 Large scale discovery of induced point mutations with high-throughput TILLING. Genome Res. 13: 524–530.

Varshney, R. K., H. Kudapa, M. Roorkiwal, M. Thudi, M. Pandey *et al.*, 2012 Advances in genetics and molecular breeding of three legume crops of semi-arid tropics using next-generation sequencing and high-throughput genotyping technologies. J. Biosci. 37: 811–820.

Wasilewska, A., F. Vlad, C. Sirichandra, Y. Redko, F. Jammes *et al.*, 2008 An update on abscisic acid signalling in plants and more... Mol. Plant 1: 198–217.

Xin, Z., M. L. Wang, N. A. Barkley, G. Burow, C. Franks *et al.*, 2008 Applying genotyping (TILLING) and phenotyping analyses to elucidate gene function in a chemically induced sorghum mutant population. BMC Plant Biol. 8: 103.

Zander, M., D. A. Patel, A. Van de Wouw, K. Lai, M. T. Lorenc *et al.*, 2013 Identifying genetic diversity of avirulence genes in *Leptosphaeria maculans* using whole genome sequencing. Funct. Integr. Genomics 13: 295–308.

Zhu, Y., H. Mang, Q. Sun, A. Hipps, and J. Hua, 2012 Gene discovery using mutagen-induced polymorphisms and deep sequencing: application to plant disease resistance. Genetics 192: 139–146.

Zuryn, S., S. L. Gras, K. Jamet, and S. Jarriault, 2010 A strategy for direct mapping and identification of mutations by whole-genome sequencing. Genetics 186: 427–430.

Communicating editor: N. D. Young

Chapter 3

Identification of mutation in an ABA insensitive *Beyma* mutant using a candidate gene approach

3.1 Abstract

Abscisic acid works in many pathways in plant biological systems, such as seed germination and plant response to stress, in which it interacts with various genes that are required in the pathways. The gene interaction provides a good opportunity in undertaking a candidate gene approach to identify a mutated gene in the ABA insensitive *Beyma* genome. A variety of events involving an ABA-gene interaction facilitated the selection of genes as candidates in this study. Hence, a number of genes were selected from ABA biosynthesis to ABA perception and signaling pathways. Preliminary analysis on 32 candidates failed to identify any unique mutation in the *Beyma* genome. Intensively, more genes were chosen as candidates to be analysed through two batches of genome sequencing of *Beyma* and WTS. As a result, eight candidate loci were found to be mutated and four of them were identified from two sequence datasets from different *Beyma* and WTS individuals. The four locus sequences putatively functions as *ABI1*, *ABI2*, *HAB1*, *HAB2*, *ABI3*, *ABI4*, and *ABSCISIC ACID 8'-HYDROXYLASE 2.* Candidates where the mutation occurred uniquely in *Beyma*; *ERD7* and *ABSCISIC ACID 8'-HYDROXYLASE 1/ P450 CYP707A1,* were predicted to cause changes in the downstream regions of the genes. Although these mutations did not affect the translated regions of the genes, they should not be omitted. Conclusively, the candidate gene approach has eliminated a number of genes as the putatively causal mutation of ABA insensitive *Beyma*.

3.2 Introduction

Sesquiterpene ABA is largely known as a key element in a wide range of plant development and responses to environmental stresses (Kermode, 2005). Its effects are varied in regulating the events of plant physiology, which require gene interaction and/or crosstalk with other hormones (as reviewed in Kermode, 2005; Fujita et al., 2006; Rock et al., 2010). This hormone also acts as a positive or negative regulator in the presence of environmental cues and during plant development (such as seed maturation, dormancy and germination, as well as seedling growth). Generally, ABA influences the seed development via concentration changes during different stages of seed maturation and dormancy breakage. However, decreased ABA concentration is not sufficient to break dormancy without the seed's acquisition of other regulatory factors such as a gibberellin signal in promoting seed germination (as reviewed in Kermode, 2005).

Numerous genes have been reported to be involved in the ABA networking system. Several genes are known to play major roles in ABA perception and signaling as described in Chapter 1. However, ABA perception and signalling are not limited to those genes only. Moreover, different sets of genes work as mediators, repressors or inducers in ABA signaling, depending on the role of ABA (Cutler et al., 2010; Kim et al., 2010; Joshi-Saha et al., 2011; Hauser et al., 2011; Ng et al., 2014). For example, ABA acts as a positive or negative regulator in interaction with genes (Rock et al., 2010) such as *AtAIB* (*Arabidopsis* basic helix-loop-helix-ype protein; Li et al., 2007), *TaCCaMK* (calcium/ calmodulin-dependent protein kinase; Yang et al., 2011), *MYB*2 (MYB transcription factor; Shan et al., 2011), *GROWTH REGULATING FACTOR 7* (*GRF7*; Kim et al., 2012) and *AREB/ABF* (bZIP transcription factors; Yoshida et al., 2014) due to environmental stresses. In the legume system, ABA is known as a negative regulator of nodulation by controlling processes required for nodule development, including bacterial infection, Nod factor signaling and consequently, nodulin gene expression (Bano and Harper, 2002; Ferguson and Mathesius, 2003; Suzuki et al., 2004; Ding et al., 2008; Tominaga et al., 2010).

With the current technology and informative resources available in plant systems, many techniques were developed to facilitate and provide rapid methods in investigating genes of interest linked to a phenotype including ABA-related traits. One of them is the candidate gene approach, which was introduced in the late 1990's for human and animal

genetics (as reviewed by Pflieger et al., 2001). A candidate gene is a gene with known function, which is involved in a metabolic pathway or influences a phenotype of interest. This approach is now being successfully applied in plant research (Zhu et al., 2012a; Patel and Patel, 2013) and other studies, such as ecological (Piertney and Webster, 2010; Smadja et al., 2012) and epidemiological studies (McQuibban et al., 2010; Landrø, 2014).

This chapter describes the application of the candidate gene approach to identify a mutation uniquely occurring in selected gene sequences of the *Beyma* genome. A number of genes were chosen directly or indirectly based on research articles including reviews, in order to obtain a list of genes putatively involved with ABA from biosynthesis to signaling pathways in plants, mainly in *A. thaliana*. Here, orthologous candidate genes of *L. Japonicus* were identified and searched for mutations in the orthologous sequences of the three genomes, namely the ABA insensitive mutant-*Beyma*, WTS of mutant and WT itself. Deep sequencing of the three genomes identified the location of mutations in *Beyma* and WTS. If a mutation occurred only in the *Beyma* sequence, it would show that the candidate with the mutation is putatively the causal gene for the *Beyma* phenotype of ABA insensitivity.

The putative causal gene will be verified to confirm it is the actual causal mutation using PCR sequencing in all three genomes and later, complementation of the mutated gene in MG-20 in order to observe the mutant phenotype and hence, validate the causal mutation. Since this project also involved the identification of unique *Beyma* mutations by undertaking genomic comparative analysis, outcomes from this chapter will be compared with the comparative outputs. Nevertheless, the candidate gene approach will assist to reduce the complexity of the genomic analysis and focus on the targeted or candidate regions of the genome. This approach will also compile all genes connected to ABA in various functions of plant systems.

3.3 Materials and methods

3.3.1 Selection of genes involved in ABA perception and signaling pathways

Initially, a total of 32 genes reported to be involved in ABA signaling were selected as candidate genes. These genes commonly are reported in the ABA-gene interaction in

ABA perception and signaling pathways (Wasilewska et al., 2008; Cutler et al., 2010; Kim et al., 2010). In order to broaden our searches, more genes were chosen as candidates, which were identified in other research papers investigating ABA roles in plant molecular biology. They were compiled and categorised according to their role in ABA-related pathways.

3.3.2 Identification of orthologs of candidate genes

Since the *A. thaliana* genome has been completely sequenced, sequences of selected candidates were obtained from The Arabidopsis Information Resources (TAIR; http://www.arabidopsis.org/) and adopted as references. The full length *A. thaliana* sequences obtained were used as a query to search for orthologs in the *G. max* genome available in the Phytozome database (http://www.phytozome.org/search.php). Due to high similarity between genomic sequences of *L. japonicus* with *G. max* compared to *A. thaliana,* the orthologs of *G. max*, rather than *A. thaliana*, were used to identify the orthologous loci of the candidates in the *L. japonicus* genome of the Kazusa database at http://www.kazusa.or.jp/lotus/blast.html. After the construction of the Legume IP database, candidate loci of *L. japonicus* were identified using the Legume IP database (http://plantgrn.noble.org/LegumeIP/). The candidate gene name or locus ID (if found in *A. thaliana*) was filled in the gene search. Candidate sequences of *L. japonicus* were then selected from output data. All orthogous candidates were selected exclusively based on these outputs. If the *A. thaliana* ID was used as a query, *L. japonicus* sequences were chosen from the ortho group (OrthoMCL) of the query, because OrthoMCL algorithm results in a lower false positive rate (Li et al., 2012).

3.3.3 Identification of unique base changes in the *Beyma* genome

Whole genome paired-end, 100 bp, short-sequence reads (>10x coverage) for three plants of *L. japonicus* MG-20 (WT, *Beyma* mutant and WTS) were generated using the Illumina Genome Analyser IIx (GAIIx) according to manufacturer's instructions. The data, representing WT (LjDIMG_03_001), *Beyma* (LjAM3_03_001) and WTS (LjAs2538_03_001), were uploaded to TAGdb (http://flora.acpfg.com.au/tagdb/; Marshall et al., 2010). The orthologs of *L. japonicus* were used to query the three available TAGdb datasets separately. FASTA files of aligning read-pairs were downloaded and re-assembled to the corresponding genomic reference regions using GeneiousPro (Drummond et al., 2011). The differences of each candidate locus in the three deep-

sequenced genomes were then analysed. This procedure was only applied for the first 32 candidate genes.

3.3.4 Identification of SNPs in candidate loci

Read mapping and SNP calling (methods in Chapter 2) identified mutations or SNPs occurring in *Beyma* and WTS, which were sequenced from a single genome (Chapter 2 and 4) and pooled genomes (Chapter 5). Orthologous candidate loci were determined if they have SNPs in their sequences based on mutations identified in both mutants as compared to WT. The effect of the SNPs predicted by SnpEff 3.0j (Cingolani et al., 2012) was also noted.

3.4 Results

3.4.1 Candidate genes in *Arabidopsis* and other plants

In this study, a total of 67 candidates were selected from different publications (Nambara and Marion-Poll, 2005; Miao et al., 2006; De Smet et al., 2006; Wasilewska et al., 2008; Cutler et al., 2010; Kim et al., 2010; Wang et al., 2011) and categorised into their roles in ABA related pathways, which are biosynthesis and catabolism, reception and signaling or post-transcriptional regulation (Table 3.1) in various plants, mainly in *A. thaliana*. Eleven candidates were involved in biosynthesis and catabolism. Nine and 47 candidates were involved in reception and signaling pathways, respectively. In biosynthesis and catabolism pathways, most candidates are involved in transferase activity, oxidation and reduction processes. There are also candidates, which catalyse hydrolysis, cleavage and/or isomerisation. Selected ABA perception candidate genes are mainly involved in serine/ threonine phosphatase activity and molecular binding. Signaling and post-transcriptional candidate genes have diverse functions, which mostly act in serine/ threonine kinase activity, transcription factor, molecular binding and transport.

3.4.2 Orthologs of candidate genes

The first 32 candidate genes were listed in Chapter 2, which was submitted as a research article for the EMS effect on the *L. japonicus* genome. As mentioned previously in Chapter 2, none of those genes were identified to be uniquely mutated in *Beyma* based on manual searches from TAGdb read datasets. Therefore, more genes were selected as

Table 3.1: Candidate genes involved in ABA related pathways. Genes were categorised into three classes with their molecular functions and loci.

Pathway	Gene	Molecular function	Gene locus
Biosynthesis and catabolism	*1-DEOXY-D-XYLULOSE 5-PHOSPHATE SYNTHASE* (DXR)	1-deoxy-D-xylulose-5-phosphate reductoisomerase activity	chr1.CM0088.390.r2.d
	1-DEOXY-D-XYLULOSE 5-PHOSPHATE SYNTHASE (DXR)	1-deoxy-D-xylulose-5-phosphate reductoisomerase activity	chr2.CM0177.10.r2.d
	1-DEOXY-D-XYLULOSE 5-PHOSPHATE SYNTHASE (DXR)	1-deoxy-D-xylulose-5-phosphate reductoisomerase activity	chr4.CM0387.120.r2.m
	1-DEOXY-D-XYLULOSE 5-PHOSPHATE SYNTHASE (DXR)	1-deoxy-D-xylulose-5-phosphate reductoisomerase activity	chr4.LjB17I07.110.r2.a
	1-DEOXY-D-XYLULOSE 5-PHOSPHATE SYNTHASE (DXR)	1-deoxy-D-xylulose-5-phosphate reductoisomerase activity	chr4.LjB17I07.130.r2.a
	1-DEOXY-D-XYLULOSE 5-PHOSPHATE SYNTHASE (DXR)	1-deoxy-D-xylulose-5-phosphate reductoisomerase activity	chr4.LjB17I07.140.r2.a
	ABSCISATE BETA-GLUCOSYLTRANSFERASE	Transferase activity	chr2.CM0028.160.r2.m
	ABSCISATE BETA-GLUCOSYLTRANSFERASE	Transferase activity	chr4.CM0227.640.r2.m
	ABSCISIC ACID 8'-HYDROXYLASE 1/ P450 CYP707A1	Oxidative degradation of ABA	chr3.CM0135.410.r2.d
	ABSCISIC ACID 8'-HYDROXYLASE 2/ P450 CYP707A2	Oxidative degradation of ABA	chr2.CM0803.690.r2.m
	ABSCISIC ALDEHYDE OXIDASE (AAO) 3	Oxidation reduction	chr2.CM0545.610.r2.d
	BETA-D-GLUCOSIDASE 1	Glycosidase/ Hydrolase activity	chr1.CM0104.2800.r2.a
	LYCOPENE BETA-CYCLASE (LYCB)	Oxidation reduction	chr6.CM0013.1810.r2.d
	MOLYBDENUM COFACTOR SULFURASE/ ABA3	Transferase activity	chr3.CM0634.640.r2.m
	NINE-CIS-EPOXYCAROTENOID DIOXYGENASE 2 (*NCED2*)/ *NCED3*	Epoxycarotenoid cleavage	chr1.CM0794.180.r2.d
	PHYTOENE SYNTHASE	Epoxycarotenoid cleavage	chr2.CM0021.2320.r2.m
	PHYTOENE SYNTHASE	Transferase activity	chr2.LjT08I01.60.r2.a
	ZEAXANTHIN EPOXIDASE (ZEP)/ ABA1	Oxidation reduction	chr3.CM0426.20.r2.a

	ZEAXANTHIN EPOXIDASE (ZEP)/ ABA1	Oxidation reduction	chr3.LjT13N17.140.r2.m
Perception	*ABSCISIC ACID INSENSITIVE 1 (ABI1)*	Protein serine/ threonine phosphatase activity	chr1.CM0133.740.r2.m
	ABSCISIC ACID INSENSITIVE 2 (ABI2)	Protein serine/ threonine phosphatase activity	chr1.CM0133.740.r2.m
	ABSCISIC ACID INSENSITIVE HOMOLOGUE (ABI8)	Transcription factor	chr1.CM0398.440.r2.a
	ABSCISIC ACID INSENSITIVE HOMOLOGUE (ABI8)	Transcription factor	chr3.CM2163.130.r2.m
	FLOWERING TIME CONTROL PROTEIN (FCA)	Nucleotide binding	chr4.CM0333.320.r2.d
	FLOWERING TIME CONTROL PROTEIN (FCA)	Nucleotide binding	chr4.CM0333.490.r2.a
	GPCR-TYPE G PROTEIN 1 (GTG1)	ABA binding	chr3.CM0127.40.r2.m
	HOMOLOGY TO ABI1 (HAB1)	Protein serine/ threonine phosphatase activity	chr1.CM0133.740.r2.m
	HOMOLOGY TO ABI2 (HAB2)	Protein serine/ threonine phosphatase activity	chr1.CM0133.740.r2.m
	PYRABACTIN RESISTANCE 1 (PYR1)/ REGULATORY COMPONENT OF ABA RECEPTOR 11 (RCAR11)	ABA binding	chr2.CM0177.730.r2.m
	PYRABACTIN RESISTANCE 1- LIKE 4(*PYL4*)/ *REGULATORY COMPONENT OF ABA RECEPTOR 10* (*RCAR10*)	ABA binding	chr3.CM0116.270.r2.m
Post-transcriptional regulation/ signaling	*3-KETOACYL-COA SYNTHASE* (KCS2)	Acyltransferase activity	chr3.CM0091.1430.r2.m
	3-KETOACYL-COA SYNTHASE (KCS2)	Acyltransferase activity	chr5.CM0071.910.r2.a
	ABA HYPERSENSITIVE 1 (ABH1)/ CAP BINDING PROTEIN (CBP80)	RNA binding protein	chr1.CM0104.530.r2.m
	ABA RESPONSIVE ELEMENT-BINDING FACTOR 2 (ABF2)/ AREB1	Protein binding	chr1.CM2113.380.r2.a
	ABI 5 BINDING PROTEIN 3 (*AFP3*)	Protein binding	chr1.CM0410.380.r2.m
	ABSCISIC ACID INSENSITIVE 3 (*ABI3*) / VIVIPOROUS 1 (VP1)	DNA binding transcription factor	chr1.CM0147.920.r2.d
	ABSCISIC ACID INSENSITIVE 4 (ABI4)	DNA binding transcription factor	chr1.CM0318.160.r2.d

	ABSCISIC ACID INSENSITIVE 5 (ABI5)	DNA binding transcription factor	chr1.CM0010.100.r2.d
	ALCOHOL DEHYDROGENASE CLASS P (ADH1)	Alcohol dehydrogenase (NAD) activity	chr1.LjT43005.120.r2.a
	ALPHA-BETA HYDROLASE	Hydrolase activity	chr2.CM0312.1250.r2.m
	ATPASE 1 (AHA1)/ OPEN STOMATA 2 (OST2)	Protein binding	chr4.CM0244.50.r2.m
	CALCIUM -DEPENDENT PROTEIN KINASE 21 (CPK21)	Protein serine/ threonine kinase activity	chr2.LjT42A12.60.r2.m
	CALCIUM -DEPENDENT PROTEIN KINASE 23 (CPK23)	Protein serine/ threonine kinase activity	chr4.CM0026.550.r2.m
	CBL-INTERACTING SERINE/ THREONINE PROTEIN KINASE 6 (CIPK6)	Serine/ threonine kinase activity	chr6.CM0037.710.r2.m
	CBL-INTERACTING SERINE/ THREONINE PROTEIN KINASE 11 (CIPK11)	Protein serine/ threonine kinase activity	chr2.CM0788.190.r2.d
	CBL-INTERACTING SERINE/THREONINE PROTEIN KINASE 15 (*CIPK15*)	Protein serine/ threonine kinase activity	chr3.LjT45I18.90.r2.d
	CONSTANS (CO)	Zinc finger protein	chr1.CM0215.30.r2.d
	EARLY RESPONSIVE TO DEHYDRATION 7 (ERD7)	Senescence/spartin	chr2.CM0272.920.r2.d
	EARLY RESPONSIVE to DEHYDRATION 15 (ERD15)	Protein binding	chr1.CM0398.350.r2.d
	ENHANCED RESPONSE to ABA 1 (ERA1)	Protein farnesylation	chr2.CM0081.550.r2.d
	ETHYLENE RESPONSE FACTOR 7 (ERF7)	DNA binding transcription factor	chr1.CM0442.310.r2.d
	ETHYLENE RESPONSE FACTOR 7 (ERF7)	DNA binding transcription factor	chr3.CM0406.340.r2.d
	FLOWERING LOCUS M (FLM)	Transcription factor	chr1.CM0104.530.r2.m
	FUSCA 3 (*FUS3*)	DNA binding transcription factor	chr1.CM0104.400.r2.a
	GATED OUTWARDLY RECTIFYING K^+ CHANNEL (GORK)	Ion transport	chr2.CM0002.560.r2.m
	GATED OUTWARDLY RECTIFYING K^+ CHANNEL (GORK)	Ion transport	chr6.CM0508.670.r2.m
	G-BOX BINDING FACTOR 3 (*GBF3*)	DNA binding transcription factor	chr1.CM0105.470.r2.d
	G-BOX BINDING FACTOR 3 (*GBF3*)	DNA binding transcription factor	chr3.CM1468.200.r2.d

	GLUTATHIONE PEROXIDASE 3 (GPX3)	Peroxidase activity	chr4.CM0004.300.r2.m
	GUANINE BINDING PROTEIN BETA 1 (AGB1)	Protein binding	chr1.CM0113.1970.r2.d
	GUANINE NUCLEOTIDE BINDING PROTEIN ALPHA-1 (*GPA1*)	Protein binding	chr5.CM0034.250.r2.m
	HIGH LEAF TEMPERATURE PROTEIN 1 (*HT1*)	Protein serine/ threonine kinase activity	chr4.CM0288.800.r2.m
	LATE EMBRYOGENESIS ABUNDANT 14 (*LEA14*)	Unknown	chr1.CM0221.110.r2.m
	LATE EMBRYOGENESIS ABUNDANT 14 (*LEA14*)	Unknown	chr5.CM0743.80.r2.m
	LIPASE CLASS 3 FAMILY PROTEIN	Hydrolase activity	chr5.CM1574.670.r2.m
	MITOGEN-ACTIVATED PROTEIN KINASE 18 (MAPKKK18)	Serine/ threonine kinase activity	chr3.CM0243.430.r2.m
	MYELOBLASTOSIS 44 (MYB44)	Transcription factor	chr5.CM0096.100.r2.m
	MYELOBLASTOSIS 101 (MYB101)	DNA/ chromatin binding	chr3.CM0243.310.r2.d
	NON-SPECIFIC LIPID-TRANSFER PROTEIN 1 (LTP)	Lipid binding	chr5.CM0200.990.r2.m
	NON-SPECIFIC LIPID-TRANSFER PROTEIN 3 (LTP3)	Lipid binding	chr3.CM1961.180.r2.m
	OPEN STOMATA 1 (*OST1*)	Protein binding	chr1.CM0016.110.r2.d
	PHOSPHOLIPASE D ALPHA 1 (*PLDα1*)	Protein binding	chr2.CM1882.150.r2.a
	PHOSPHOLIPASE D ALPHA 1 (*PLDα1*)	Protein binding	chr3.CM0142.570.r2.d
	Pi TRANSPORTER (*Pht*)	Transmembrane transport	chr1.CM0295.140.r2.m
	POTASSIUM CHANNEL IN A.THALIANA 1 (*KAT1*)	Ion transport	chr6.CM1757.280.r2.a
	REGULATOR OF G-PROTEIN SIGNALING 1 (*RGS1*)	Protein binding	chr6.LjT45M05.110.r2.d
	RESPIROTARY BURST OXIDASE F (RBOHF)	Peroxidase activity	chr6.CM0013.510.r2.m
	RESPONSIVE TO DESICCATION 26 (*RD26*)	DNA binding	chr3.CM0590.350.r2.d
	SLAC1 HOMOLOGUS-1(SLAH1)	Membrane transport	chr3.CM0243.420.r2.m
	STELAR K+ OUTWARD RECTIFIER (SKOR)	Ion transport	chr6.CM0508.670.r2.m
	SUCROSE-PHOSPHATE SYNTHASE 1 (SPS1)	Sucrose-phosphate synthase activity	chr3.CM0047.470.r2.d
	SUCROSE-PHOSPHATE SYNTHASE 1 (SPS1)	Sucrose-phosphate synthase activity	chr4.CM0003.1230.r2.m
	U-BOX DOMAIN CONTAINING PROTEIN 19	Ubiquitin-protein transferase activity	chr4.CM0414.490.r2.a

candidates to identify putative causal mutation. Seventy one loci representing orthologous sequences of 67 candidates were analysed in this study (Table 3.1). The orthologous sequences ranged from 250 to 5000 bp in length. Twenty of them were located in chromosome 1 while chromosomes 2, 3 and 4 had fourteen, sixteen and ten candidate sequences, respectively. Only six and five candidate sequences were located in chromosomes 5 and 6, respectively. These orthologous loci were selected based on the output from searches in the Legume IP database. Some of the candidates had more than one orthologs in the *L. japonicus* genome and they were included in the list (Table 3.1).

3.4.3 Mutation in candidate genes

Preliminary analysis from the realignment of TAGdb results of the 32 candidates indicated an absence of mutation in these genes. Later, a new test was performed after obtaining a new list of SNPs from the third data analysis method to ensure the results were reliable and accurate. This test identified the presence of SNPs in either *Beyma* or both *Beyma* and WTS as compared to WT. From the sequencing of the single genome of *Beyma* and WTS, eight candidates had shown to have SNP(s) in their sequences either in one of the mutants or both (Table 3.2). *ABI3* gene at the locus named chr1.CM0147.920.r2.d was mutated in WTS but not in *Beyma*. Meanwhile, two candidates were mutated uniquely in *Beyma*. They were *ERD7* (chr2.CM0272.920.r2.d) and *ABSCISIC ACID 8'-HYDROXYLASE 1/ P450 CYP707A1* (chr3.CM0135.410.r2.d). Two SNPs were located in both mutants at the same candidate locus named chr2.CM0803.690.r2.m, which encodes for *ABSCISIC ACID 8'-HYDROXYLASE 2/ P450 CYP707A2*. The other SNPs were present in both *Beyma* and WTS. In addition, all of these SNPs were predicted to cause mutations in the upstream or downstream regions of the corresponding genes in the MG-20 genome, hence, they were not included in the final list of putative causative SNPs in Chapter 4.

Re-sequencing of *Beyma* and WTS from pooled genomes resulted in a different output (Chapter 5). Candidate gene analysis was also carried out by looking at existence of mutations in our candidates based on the re-sequencing data (Table 3.3). Four mutations were listed as present at the same positions as the previous candidate locus (Table 3.2). These mutated sequences hit to *ABI1*, *ABI2*, *HAB1*, *HAB2*, *ABI3*, *ABI4*, and *ABSCISIC ACID 8'-HYDROXYLASE 2/ P450 CYP707A2* genes. *ABI 5 BINDING PROTEIN 3* gene locus named chr1.CM0410.380.r2.m had a mutation, which was predicted to change amino acid nonsynonymously. However, this mutation only occurred

Table 3.2: Mutations identified in candidate loci based on the sequencing data of the single genome of *Beyma* and WTS.

Gene locus	Gene identity	Ref	Change	Effect	Mutation	
					Beyma	**WTS**
chr1.CM0133.740.r2.m	*ABI1*	C	T	Upstream	Yes	Yes
chr1.CM0147.920.r2.d	*ABI3*	T	G	Downstream	No	Yes
chr1.CM0215.30.r2.d	*CO*	T	C	Upstream	Yes	Yes
chr1.CM0318.160.r2.d	*ABI4*	T	C	Upstream	Yes	Yes
chr2.CM0272.920.r2.d	*ERD7*	A	G	Downstream	Yes	No
chr2.CM0788.190.r2.d	*CIPK15*	T	C	Upstream	Yes	Yes
chr2.CM0803.690.r2.m*	*P450 CYP707A2*	G	A	Upstream	Yes	Yes
chr2.CM0803.690.r2.m*	*P450 CYP707A2*	G	A	Upstream	Yes	Yes
chr3.CM0135.410.r2.d	*P450 CYP707A1*	T	C	Downstream	Yes	No

*Different positions; Ref: reference.

Table 3.3: Mutations identified in candidate loci based on the re-sequencing data of the pooled genomes of *Beyma* and WTS.

Gene locus	Gene identity	Ref	Change	Effect	Mutation	
					Beyma	**WTS**
chr1.CM0133.740.r2.m**	*ABI1*	C	T	Upstream	Yes	Yes
chr1.CM0147.920.r2.d**	*ABI3*	T	G	Downstream	Yes	Yes
chr1.CM0318.160.r2.d**	*ABI4*	T	C	Upstream	Yes	Yes
chr1.CM0410.380.r2.m	*AFP3*	C	T	Nonsynonymous coding	No	Yes
chr1.CM0794.180.r2.d	*NCED2*	T	G	Intragenic	Yes	Yes
chr2.CM0545.610.r2.d	*AAO3*	C	A	Downstream	No	Yes
chr2.CM0803.690.r2.m**	*P450 CYP707A2*	G	A	Upstream	Yes	Yes
chr6.CM1757.280.r2.a	*KAT1*	C	A	Downstream	No	Yes

**The same candidates as listed in Table 2; Ref: reference.

in WTS. An intragenic effect was also predicted from a mutation in *NCED2* gene locus named chr1.CM0794.180.r2.d in both *Beyma* and WTS genomes.

3.5 Discussion

The selection of candidate genes was initially based on *Beyma* phenotype, which was easily dehydrated and showed insensitivity to ABA (Biswas et al., 2009). A highly similar phenotype to *Beyma* was found in *A. thaliana* that mutated in *ABI1* gene (Merlot et al., 2001). However, a *L. japonicus* ortholog of *AtABI1* was not altered in both MG-20 WT and *Beyma* (Biswas et al., 2009). Fujii and Zhu (2009) also reported similar phenotype on triple mutation of protein kinases (*OST1*, *SnRK2*.2 and *SnRK2*.3) in *A. thaliana*, in which ABA-inhibition of seed germination was reduced and susceptibility of dehydration was increased. These mutated genes were reported playing roles in ABA perception and signaling pathways, suggesting that *Beyma* might be defective in these pathways. Thus, candidates were chosen from these pathways to identify the causal mutation. However, more candidates were nominated later due to the absence of mutation in the first 32 candidates. The broad selection gave higher chances to identify the causal mutated gene. Table 3.1 was made to demonstrate the role of each candidate with their molecular function in ABA related pathways (www.uniprot.org). This list could be used as a reference for future work on ABA related analyses in *L. japonicus* and other plants.

The first method of candidate gene analysis was the application of TAGdb searches (Marshall et al., 2010) to identify uniquely mutated sequence in this study. This method was time consuming and produced false positive result. During BLAST searches, locus sequence (less than 5000 bp in length) was put as a query to obtain paired reads of the genomes. As a result, paired reads that aligned to the query sequence might belong to different positions of the genome. The whole-genome duplication of *L. japonicus* showed that 13 % of genes assigned in six large duplicated regions between chromosomes were conserved in each pair of duplicated segments (Sato et al., 2008). This affected the read alignment of short locus sequence in the TAGdb BLAST and therefore, caused the alignment of non-allelic homologous reads to locus sequence during the re-assembly in this study.

After read mapping and SNP calling (Li et al., 2009; Lorenc et al., 2012) were carried out to obtain SNPs occurring in the three genomes, the second method was performed. The identification of SNPs facilitated the analysis of candidate sequences in this study. Candidate sequences of *L. japonicus* were obtained based on their orthologs in *A. thaliana* and other plants such as *G. max* and *M. truncatula* (http://plantgrn.noble.org/LegumeIP/search.do). The determined genome sequence of *L. japonicus* MG-20 was assembled covering 67 % of the whole length with 91.3 % of gene space annotated (Sato et al., 2008). The sequence data allowed the recognition of sequences or loci in the *L. japonicus* genome for majority of genes with known sequences in other plants. Therefore, the selection of candidate sequences is not a difficult task in this study.

In this project, two batches of genome sequencing were performed to intensify the identification of the causal mutated gene (Chapters 4 and 5). The sequencing data were used to analyse the presence of mutations in our candidate genes. Although two candidates, *ERD7* and *ABSCISIC ACID 8'-HYDROXYLASE 1/ P450 CYP707A1*, were found to be mutated uniquely in *Beyma*, they were not short listed as potential causative mutated genes in Chapter 4. The identified mutations were located in the downstream regions of the corresponding genes. Previously, *P450 CYP707A1* gene had been proved to participate in ABA catabolism inside guard cells during high humidity condition (Okamoto et al., 2009). Moreover, *ABI1*, *ABI2*, *HAB1*, *HAB2*, *ABI3* and *ABI4* are also known as important elements in ABA signaling (reviewed by Nakashima and Yamaguchi-Shinozaki, 2013). These show that *CYP707A1* and other genes are good causal candidates in *Beyma*. Thus, these candidates should not be omitted and could be verified later if none of the potential causal SNPs is the actual causative mutation. However, this candidate gene approach could eliminate the candidates as a putative causal *Beyma* gene.

In addition, this analysis also showed the effects of EMS on our candidate sequences based on the identification of mutations throughout the genome of both mutants. It showed the effect of EMS on the ABA candidates might reflect other phenotypes in *Beyma* or/ and WTS, providing clues on reverse genetics analysis in *L. japonicus*. We also suggest that EMS gave impacts in small parts of ABA linked pathways, where only ~12 % of the ABA genes were impaired in our mutants. Although these mutations might not be the *Beyma* gene, they were determined as background mutations

due to EMS mutagenesis. These could be interest of mutational analyses in ABA related pathways of *L. japonicus* and other legumes.

Nevertheless, the candidate gene approach showed the possibility of ABA insensitive *Beyma* mutation could be in a gene, which is yet not associated with ABA biosynthesis or signaling. In addition, *Beyma* was previously isolated from a segregation analysis of a single gene (monogenic trait; Biswas et al., 2009), which increase the rejection of candidates. If time permitted, candidate selection could also be done on genes associated in other hormones such as gibberrelin and ethylene, which are antagonist of ABA in seed dormancy (Kermode, 2005) and root growth (Steffens et al., 2006), respectively. Like other hormones, ABA requires an intertissue transport system that involves a set of genes to transfer ABA from vascular to guard cells (Cutler et al., 2010; Kim et al., 2010; Kuromori et al., 2014). These interconnection pathways indicate how ABA works extensively in plant systems, demonstrating the challenge in the candidate gene approach to identify the *Beyma* gene. A different approach was therefore followed in our search for the causal gene. Called SNPs were further analysed and the output data are discussed in the next chapters.

3.6 Conclusion

Many genes interact with ABA directly or indirectly due to the presence of environmental cues or developmental processes. This gives a good opportunity in the selection of genes as candidates in this study. However, the identification of putative causal mutation is very challenging. Although 71 loci that represent 67 genes were nominated, but none of them showed a unique *Beyma* mutation with nonsynonoymous changes in the annotated genes of MG-20. Thus, SNP analysis had to be performed.

Chapter 4
Identification of a potentially causative mutation in the *Beyma* mutant

4.1 Abstract

Different technologies can be applied to identify a gene of interest which has been impaired due to mutagenic treatment. With the rapid development of second generation sequencing technology, many tools have been developed to accomplish the objective of mutagenesis studies. Here, this technology was employed to identify a causative gene in an ABA insensitive *Beyma* mutant of the model legume *L. japonicus*. Whole genome sequencing was performed on a single plant of *Beyma*, a WTS of the mutant and a WT *L. japonicus*. These genome sequences were subjected to comparative genomic analysis by looking at SNPs between them. The objective was to identify a causative mutated gene in the ABA insensitive *Beyma* genome. The causative gene should contain a mutation, which is only present in the *Beyma* genome but absent in WTS and WT. In preliminary analyses, a number of candidate genes were identified with putatively causative mutations; however the candidates were later verified as background mutations. Nevertheless, a new list of genes or loci was subsequently identified to be putatively causative mutations and will be verified. This chapter also demonstrated the selection of an F2 population (outcross of *Beyma* and *L. japoncus* ecotype Gifu) in order to analyse the segregation of a putative causative SNP in the F2 plants carrying homozygous mutated alleles. The identification of the causative mutated gene will show a potentially novel gene that is involved in the ABA signaling pathway in legume systems.

4.2 Introduction

Forward genetics are implemented to identify genes responsible for plant growth and development. The approach typically begins with an induced mutagenesis which involves chemical, irradiation or insertional approaches to isolate mutants with desired phenotypes or beneficial physiological responses as well as to identify protein function in plant systems (Kim et al., 2006; Maple and Møller, 2007; Weil and Monde, 2009). The most common approach is the application of chemicals such as EMS, which tends to generate point mutations resulting in mis-pairing and base changes (Krieg, 1963; Sikora et al., 2011), as described in Chapter 2. This study was done on the ABA insensitive *Beyma* mutant, which was previously isolated from EMS mutagenised *L. japonicus* ecotype MG-20 seeds (Biswas et al., 2009). In order to identify the EMS derived mutation that was responsible for the phenotype of ABA insensitivity in *Beyma*, SGS technology was applied.

At present, genomic sequencing can be performed at an affordable cost (Thudi et al., 2012; Pabinger et al., 2013). However, the main concerns are; which genome will be the input for sequencing and how the sequencing output will be processed. These depend on the study goals. Various approaches have been applied to find SNPs or mutation induced by EMS. In a bulked segregant analysis, genomic DNA of a backcrossed segregant population was pooled for sequencing to increase SNP frequency and identify mutated regions (Ashelford et al., 2011; Mokry et al., 2011; Hartwig et al., 2012). Lindner et al. (2012) demonstrated a method to identify a mutated gene in a gametophyte lethal mutation. They backcrossed an EMS mutagenised plant to the non-mutagenised parent twice and identified mutated causative SNPs by measuring a SNP ratio within the *A. thaliana* genome. All these techniques require backcrossing between the isogenic parental line and EMS-induced mutant, which can be time consuming.

On the other hand, sequencing more than one individual mutant line for comparative genomic analyses between the mutants and different accessions has also been implemented to identify the causal mutation (Uchida et al., 2011). With the current information on gene and protein functions available, candidate genes and/or regions can be selected or targeted based on the mutant phenotype (Hartwig et al., 2012; Zhu et al., 2012a). However, targeting candidate genes could be inaccurate and consequently, prediction of different genes is necessary. In this study, a candidate gene approach had

been undertaken (Chapter 3). Although it failed to identify the mutated gene of *Beyma*, it helped to eliminate a number of genes as the putative causative mutation of *Beyma*.

Therefore, we present a new approach in identifying SNPs or indels (insertion and deletions) in *Beyma* using sequencing technology. We took advantage of the heterozygosity of the original dominant *Beyma* mutation to remove background damage and identify a causal mutation caused by the EMS mutagenesis (Figure 4.1). *Beyma* mutant with heterozygous alleles (Bb) were allowed to regenerate into mutants with homozygous mutant alleles (BB), heterozygous mutant alleles (Bb) and homozygous wild type alleles (bb; hence called as WTS of mutant). Instead of backcrossing the homozygous mutant with the isogenic parental line (MG-20) to subtract background mutation, we performed a comparative analysis between the homozygous *Beyma* mutant, MG-20 WT and WTS of the mutant. If SNPs or base changes present only in *Beyma*, they were putatively the causative mutation in the ABA insensitive mutant.

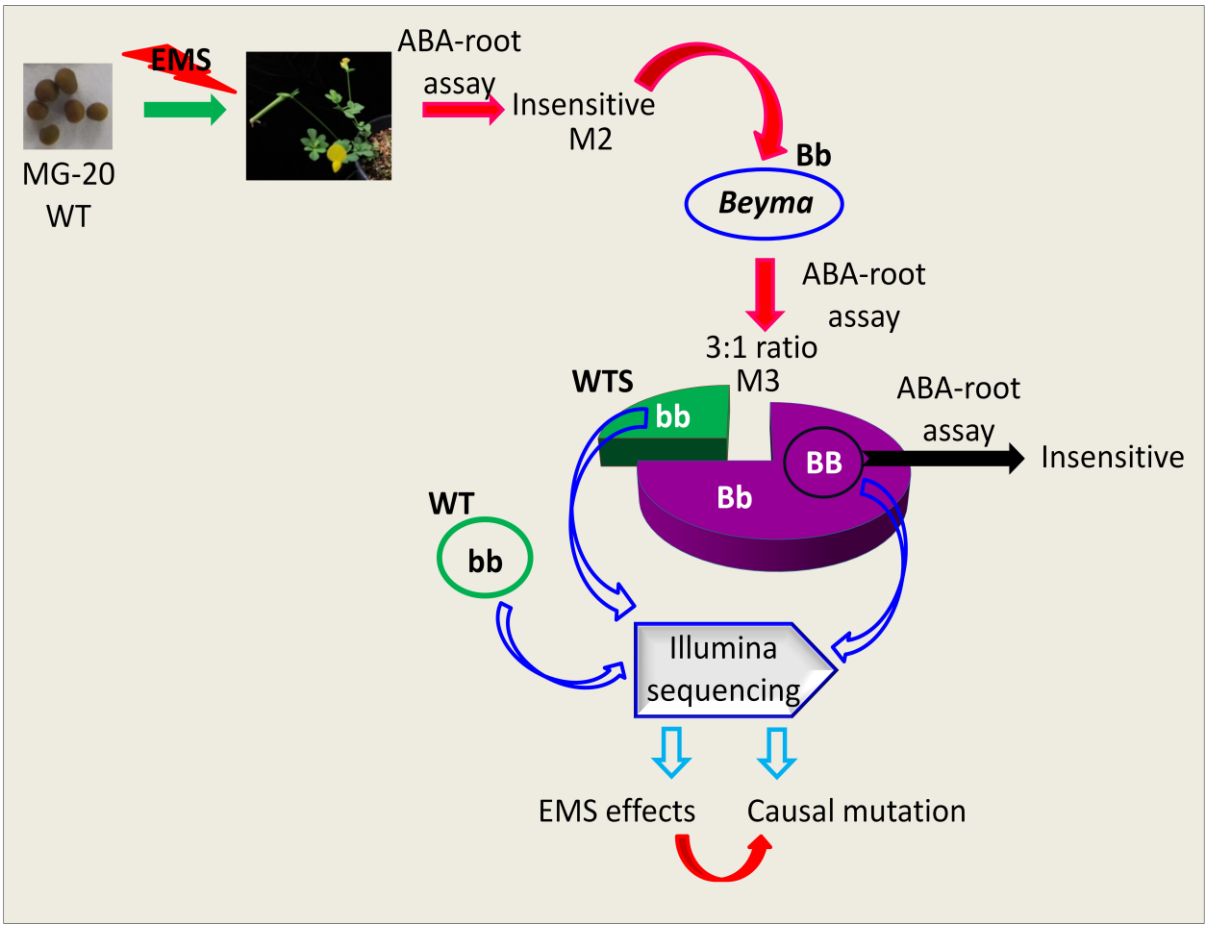

Figure 4.1: Isolation of *Beyma* mutant from a population of EMS mutagenised MG-20 WT seeds. Seeds collected from ABA insensitive M2 plants were subjected to ABA-root assay resulting in the isolation of homozygous *Beyma* (BB), heterozygous mutant (Bb) and

homozygous WTS of mutant (bb). Whole genome sequencing was performed on three individuals (WT, BB and bb) to identify the candidate causal mutation in *Beyma*.

4.3 Materials and methods

4.3.1 Plant materials

Seeds of MG-20 and its derivatives were scarified, sterilised and grown as described in Biswas et al. (2009). Five lines of *Beyma* seeds (Line B1 – Line B5), derived from the fourth generation of the originally selected homozygous *Beyma* mutant (M4) were treated with the plant hormone ABA exogenously. For the selection of homozygous *Beyma* lines, two different treatments were performed on ABA sensitivity: seed germination and seedling growth. For ABA treatment during seed germination, 100 µM of filter-sterilised abscisic acid (ABA, Sigma-Aldrich, St. Louis, Missouri) was used to wet the filter papers, whereas sterile water was used for control. The percentage of seeds that germinated was determined. Two days after germination, the seedlings were transferred to Phytatray II (Sigma-Aldrich) containing autoclaved half strength B5 medium (commercially available B5 salts and B5 vitamins (Sigma-Aldrich), 0.06 % (w/v) of MES (Sigma-Aldrich) and 1 % (w/v) of agar (Sigma-Aldrich) at pH 5.7) and grown in a CMP4030 Conviron incubator (Winnipeg, Manitoba) at 12/12 hour and 21 ^{o}C/18 ^{o}C temperature cycle of day/night for 7 days. For ABA treatment during seedling growth, water-germinated seedlings were transferred into Phytatray II (Sigma-Aldrich) containing the same B5 medium supplemented with 50 µM ABA. After 7 days of growth, the seedlings were transferred into a container with sterile water overnight (allowing the seedlings to adjust to the humidity changes) before placing them in pots containing medium-grade vermiculite. The seedlings were grown in a Conviron growth chamber at 16/8 h day/night, 24 ^{o}C/20 ^{o}C day/night temperatures and 70 % humidity (watered control).

4.3.2 Genomic DNA extraction

Genomic DNA was isolated from plant tissues using the CTAB method and subjected to RNase treatment as described in Chapter 2.

4.3.3 Dehydration screening

Outcrossing of *Beyma* pollens to stigma of *L. japonicus* ecotype Gifu was carried out previously by Dr Bandana Biswas from the Centre for Integrative Legume Research (CILR, Brisbane, Queensland). Seeds of the outcrossed F2 population were germinated and grown as described above. The seedlings were watered with half-strength B5 liquid medium on alternate days and the plants were screened after 5 weeks. The whole shoot part of each F2 plant (above the vermiculite) was cut and the second trifoliate leaflet was removed for separate observation. They were kept at room temperature for 2 hours before visually scoring the leaflets for dryness (Biswas et al., 2009). The leaflets/shoots susceptible to dryness were scored as F2 plants with mutant phenotype and quick frozen in liquid nitrogen and stored at -80^{o}C for further analysis. Dehydration screening was also performed on *Beyma*, MG-20 and Gifu as controls. F2 individuals showing a positive result in both tests were selected as putative mutated locus carriers.

4.3.4 Genomic sequencing

Deep sequencing of the whole genome was performed for three single plant genomes of MG-20 (WT, *Beyma* and WTS) as described in Chapter 2.

4.3.5 Read mapping and SNP calling

Three procedures were performed in this study to obtain a good list of candidate SNPs. The genomic sequence of MG-20 WT was downloaded from the Kazusa database, Miyakogusa.jp 25 (www.kazusa.or.jp/lotus/) and used as a reference. Procedures 1 and 2 (Figure 4.2) were performed by Dr Stephen Kazakoff from the Queensland Centre for Medical Genomics (QCMG, Brisbane, Queensland). Procedure 3 was carried out by Mr Pradeep Ruperao from the ACPFG.

4.3.5.1 Procedure 1

All paired reads of three deep-sequenced genomes (WT, *Beyma* and WTS) were separately mapped to the concatenated contigs of the reference with default parameters using the programs Burrows Wheeler Aligner (BWA; Li and Durbin, 2009) and Samtools (Li et al., 2009). SNPs were then called from the three aligned sequences as compared to the reference using FreeBayes (http://bioinformatics.bc.edu/marthlab/FreeBayes). Custom Perl scripts were written to compile all SNPs in WT and WTS prior to subtraction of a copy of common SNPs in both genomes. The filtration was processed using Perl and/or UNIX commands to remove common SNPs in all the three genomes and to retain SNPs with; (i)

EMS-canonical G/C-to-A/T base substitution, (ii) alternate base count occurring three or more times and (iii) reference base count occurring zero time.

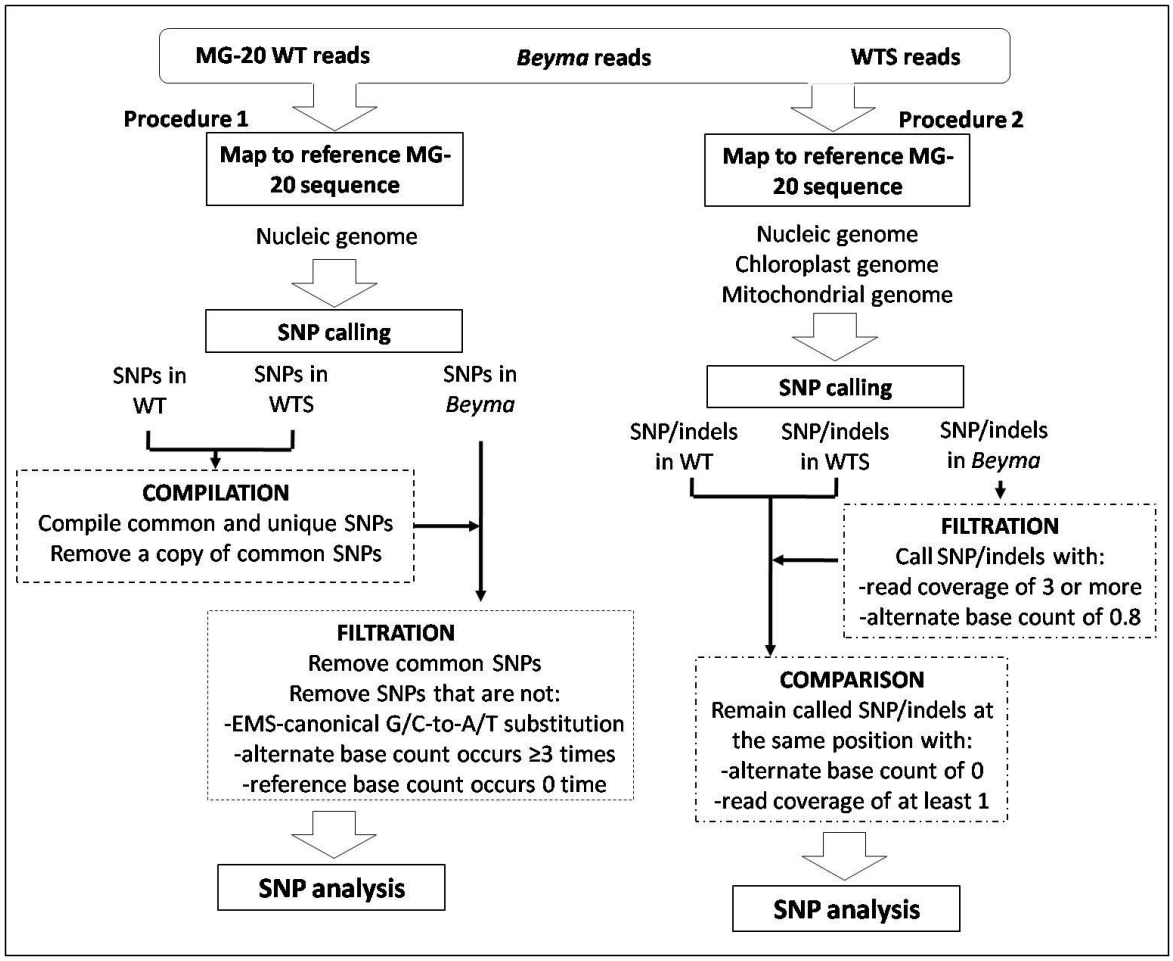

Figure 4.2: Pipeline of two methods performed in this study. Each method was explained in the text.

4.3.5.2 Procedure 2

Procedure 2 was established after the assembly of the *L. japonicus* mitochondrial genome, which was performed by Kazakoff et al. (2012). This method included nucleic, mitochondrial and chloroplast genomes for the mapping procedure (Kato et al., 2000; Figure 4.2). This step was assigned to allow all paired reads to map against their specific regions. All the contigs of reference sequences were not concatenated in this method. The paired reads of each genome were mapped to the reference sequences using BWA. Custom Perl scripts were written to call SNPs and indels that occurred in *Beyma* using the Bio:DB:Sam modules (http://search.cpan.org/dist/Bio-SamTools/). The mapping was visualised using the program Tablet (Milne et al., 2010). We called SNPs and indels with allele balance ratio of alternate base of 0.8 and the read coverage of 3 or more. The

SNPs/indels were screened by comparing with the read alignments of WT and WTS genomes. The SNPs/indels occurring at the same position in either one or both WT and WTS alignments (with allele balance ratio of alternate base of 0 and the read coverage of at least 1) were removed to subtract natural variation and background mutation. The SNP/indel numbers were scaled down by selecting alteration that occurred only in the coding region.

4.3.5.3 Procedure 3

Read mapping and SNP calling were described in Chapter 2 under the subtopic 'Sequencing and SNP Identification'.

4.3.6 SNP analysis

After the identification of SNPs, further analysis was employed to identify a putative causal mutated gene.

4.3.6.1 Procedure 1

Two hundred bp flanking each SNP at both ends were extracted from the reference sequence producing a number of 400-bp fragments. These fragments were used to query the Kazusa database, using BLASTN. The loci which obtained a hit (hence called putative SNP loci) were selected to analyse whether the SNP was located in the coding region or not. The putative protein functions and possible amino acid changes (synonymous and nonsynonymous mutations) were determined. These procedures were performed by Dr Stephen Kazakoff from the QCMG. The putative SNP loci with the nonsynonymous changes were further analysed by searching for ABA-related articles that documented the correlation of those proteins with ABA directly or indirectly. All of the putative SNP loci were also used to query TAGdb, as a candidate gene approach. The occurrence of mutations between the three genomes was also assessed. The putative SNP loci with variant base (occurs only in *Beyma*) were amplified and sequenced in *Beyma* and WT to confirm the mutation.

4.3.6.2 Procedure 2

Custom Perl scripts were written to identify SNPs/indels which were located in exons (based on the annotated Kazusa model file), translate each codon into its relevant amino acid and print the complement sequence of coding region and the alignment of protein sequences between WT and *Beyma*. These procedures were performed by Dr

Stephen Kazakoff from the QCMG. The procedures facilitated subsequent sequence analysis and primer design for the confirmation of mutation using amplification and sequencing in *Beyma* and WT, as in procedure 1. If the mutation was only present in *Beyma*, the locus would be sequenced in WTS. This step was to ensure the mutation was not a background mutation that caused by the EMS mutagenesis.

4.3.6.3 Procedure 3

SNPs occurring uniquely in *Beyma* were subtracted from the list and analysed for nonsynonymous and synonymous changes. These changes were predicted using SnpEff 3.0j (Cingolani et al., 2012) according to their effect on MG-20 annotated genes (Sato et al., 2008).

4.3.7 PCR sequencing

Forward and reverse primers were designed, based on the extracted 400 bp fragments using Primer3Plus (http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi) and synthesised by Sigma-Aldrich. A final volume of 50 µL PCR mix was prepared as follows; 1X PCR buffer (Scientifix, Melbourne, Victoria), 0.25 mM dNTPs, 0.5 µM of each primer, 2.5 U of Taq DNA polymerase (Scientifix), 30-50 ng of genomic DNA and 33.5 µL of sterile water. Amplification was then carried out as follows: 94 ^{o}C for 5 min, 35 cycles of 94 ^{o}C for 45 s, 60 ^{o}C for 30 s, and 72 ^{o}C for 2 min followed by extension at 72 ^{o}C for 10 min. The amplified products were analysed on a 2 % (w/v) TAE agarose gel and purified using the QIAquick PCR purification kit (Qiagen, Hilden, Germany) according to manufacturer's instruction. The purified amplicons were sequenced by the Australian Genome Research Facility (AGRF, Brisbane). The sequencing results were analysed using Geneious Pro 5.1.5 (Drummond et al., 2011).

4.4 Results

4.4.1 Selection of homozygous *Beyma* lines

Table 4.1 shows the number of seeds and seedlings of MG-20 WT and five *Beyma* lines upon ABA treatment during germination and root growth. MG-20 WT seeds did not germinate in ABA but had 75 % (15/20) germination without ABA. Lines B1 and B2 showed a low rate of germination with or without ABA. A good germination rate was

observed in lines B3, B4 and B5 for both treatments, displaying their ABA insensitivity. All MG-20 seedling roots had shorter length. Meanwhile, more than 75 % of line B3 and B4 produced longer roots in a B5 medium supplemented with ABA compared to MG-20 WT. Due to a small number of samples; root growth of germinated seeds should be done in the absence of ABA as well, as control. However, germination results demonstrated that all *Beyma* lines are homozygous population. But, lines B3 and B4 were good lines to be used as homozygous *Beyma* mutants for subsequent analyses in this study because they have a good germination rate. This result was supported with a different test which will be described in Chapter 5.

Table 4.1: Outcomes from ABA treatment on MG-20 WT and *Beyma* lines.

Line	Germination				Root growth (>5mm)*
	No. of seeds	No ABA	No. of seeds	With ABA	With ABA
WT	20	15	25	0	0
B1	19	7	25	14	3
B2	20	8	24	6	4
B3	20	18	24	24	9
B4	20	15	24	24	10
B5	20	13	25	19	7

*taken from water-germinated seeds

4.4.2 Screening of F2 population

Dehydration effect on leaflets was recorded 2 hours after plucking, by which time the difference was noticeable (Figure 4.3A-D). In this study, leaflets of MG-20, Gifu and *Beyma* responded as expected (Biswas et al., 2009). In the shoot drying test, MG-20 and Gifu plants did not dehydrate (Figure 4.4A-D). All tested *Beyma* plants dehydrated after 2 hours (Figure 4.4E-F). Among 209 F2 plants, 37 % (78 individuals) of F2 leaflets shrunk and 44 % (92 individuals) of shoots dehydrated. These results determined a ratio of dehydrated F2 to not dehydrated F2 plants was 1:1.68 (78:131) and 1:1.27 (92:117) for leaflet and shoot drying tests, respectively. Meanwhile, only 50 individuals dehydrated in both tests, in which a ratio of 1:3.18 for dehydrated to not dehydrated F2 was observed. Only these 50 individuals were selected for subsequent analysis.

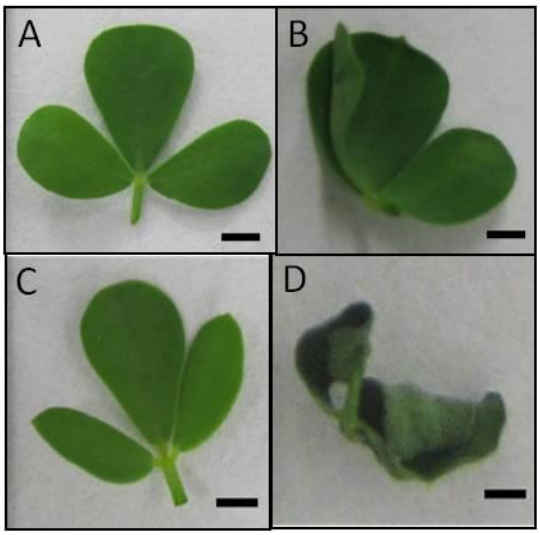

Figure 4.3: Detached trifoliate leaflets of a F2 plant showed the dehydration response after 2-hour treatment. A WT leaflet displayed less dehydration after two hours (A, before; B, after). *Beyma* leaflet was considerably more dehydrated after the 2-hour test (C, before; D, after). Bars represent 0.2 cm.

4.4.3 Putative SNPs in the *Beyma* genome

4.4.3.1 Procedure 1

WT reads showed 13,489,648 putative SNPs compared to the Kazusa reference; whilst WTS and *Beyma* have 11,756,284 and 14,797,526 putative SNPs, respectively (Figure 4.5A). The selection of EMS canonical base mutation decreased the number of putative SNPs to 33,605 in which 49.9 % (16,760) and 50.1 % (16,485) were C-to-T and G-to-A base changes, respectively. This number was reduced by removal of heterozygous SNPs, in which the mapped reads contained a reference base. The removal step identified only 0.6 % of putative SNPs (101 C-to-T and 100 G-to-A base substitutions) in a 618,226,588 bp concatenated sequence for further analysis. Twelve out of 201 SNPs (Table 4.2) showed consequent nonsynonymous changes in which only two SNPs (SNP2 and SNP5) led to a termination of translation. Table 4.3 shows the details about each putative SNP. All putative SNPs had a total of three alternate bases (AB) or reads mapped except SNP2 and SNP6 (four AB). SNP7 showed a relatively higher number of AB (which was 8).

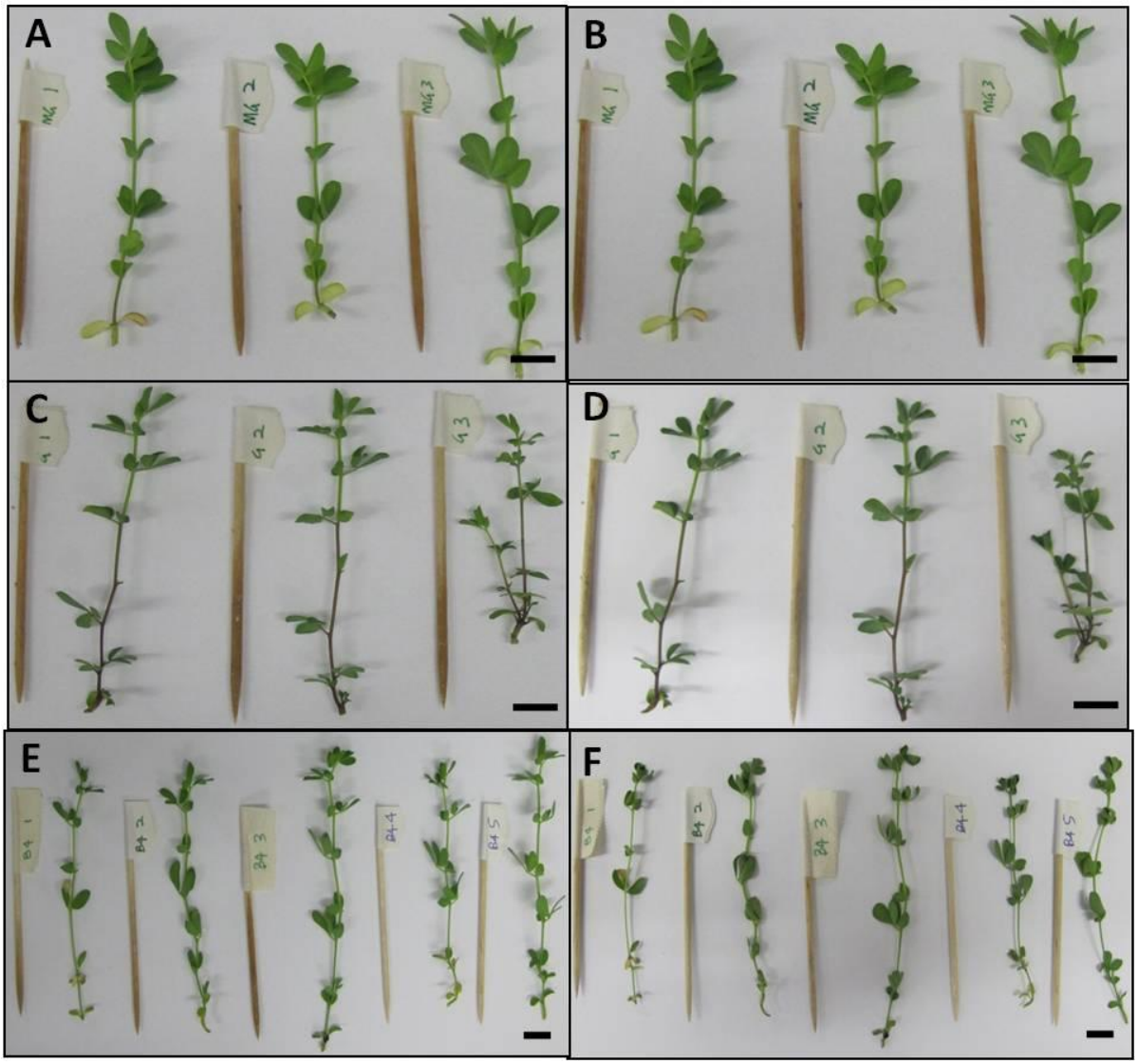

Figure 4.4: Three 5-week-old seedlings of MG-20, Gifu and *Beyma* before and after the shoot drying test. Seedlings of MG-20 (MG) and Gifu (G) showed no difference before (A and C) and after (B and D) the drying test, respectively. However, seedlings of *Beyma* showed the dryness effect due to the drying test (E, before; vs. F, after). Bars represent 1 cm.

Further analysis was carried out by repeating a similar approach as the candidate genes. The BLAST analysis for each candidate locus with a SNP was performed in TAGdb and their reads were realigned to the locus sequence. Only SNP6 and SNP10 showed the presence of a variant base only in *Beyma* but not in WT or WTS (Table 4.3). Three of the candidate loci; SNP5, SNP6, and SNP9 encoded for proteins that have been reported to be involved in or correlated with the ABA signaling pathway. The SNP5 locus encodes a small glutamine-rich tetratricopeptide repeat-containing protein that showed mRNA accumulation in *A. thaliana* seedlings in response to ABA treatment (Clément et al., 2011).

The SNP6 locus encodes a plasma membrane receptor-like kinase wherein a mutation in a receptor protein kinase (RPK1) of *A. thaliana* displayed insensitive behaviour towards ABA (Osakabe et al., 2005). The homeodomain-like protein putatively coded by SNP9 locus was found to be correlated to ABA signaling regulation as reported by Himmelbach et al. (2002) and Son et al. (2010). However, only SNP6 and SNP10 were found to be mutated in the *Beyma*.

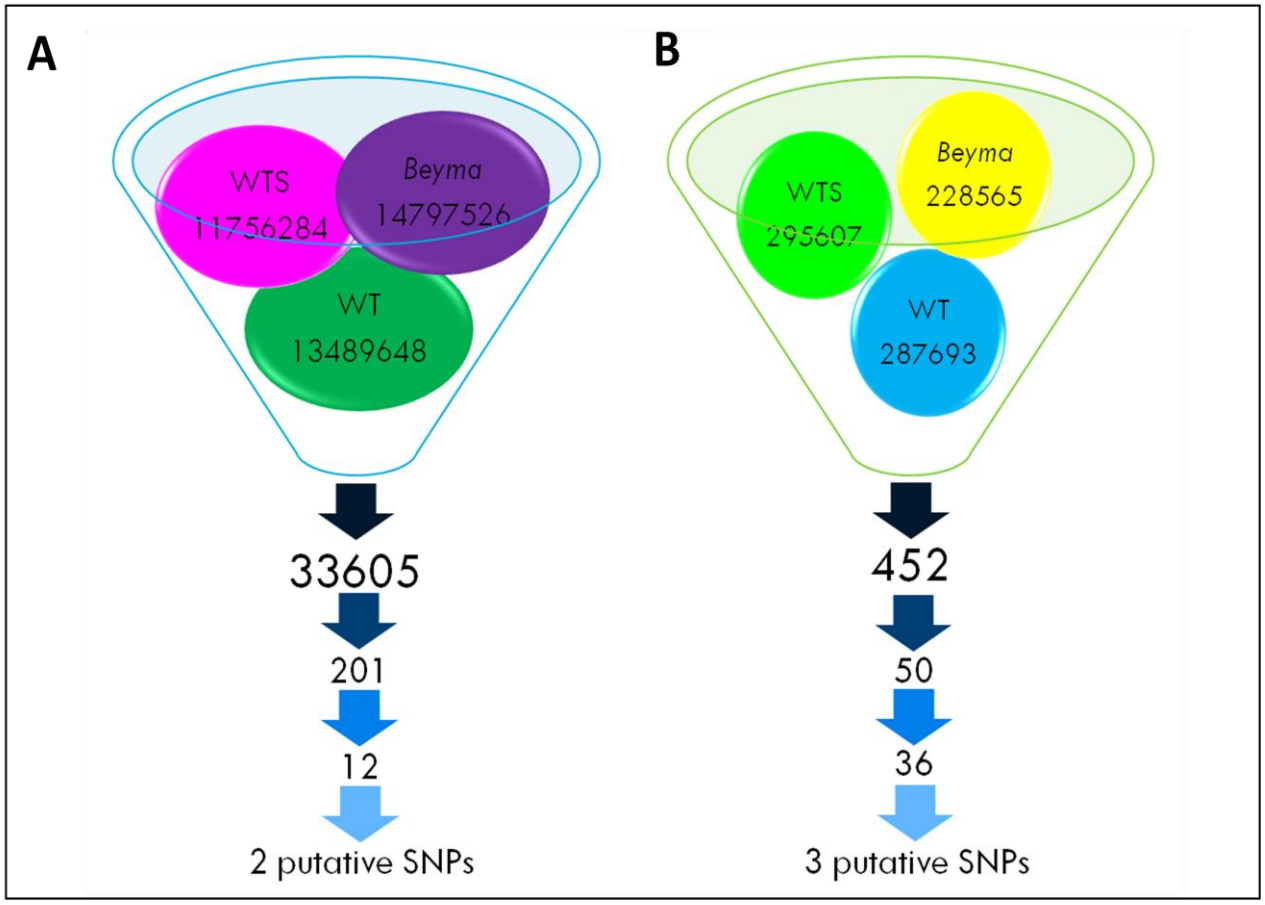

Figure 4.5: Total number of putative SNPs and/or indels after SNP calling and filtration were carried out using procedures 1 (A) and 2 (B). SNP calling produced a high number of putative SNPs extracted from each genome and further filtration resulted in the reduction of the total number of putative SNPs to two (A) and three (B).

Table 4.2: Putative SNPs occurring in the *Beyma* with their loci in the *Lotus* genome, putative function and amino acid changes (from procedure 1).

SNP name	Kazusa database				NCBI blastx					Amino acid change
	Locus	Putative function	Score	E value	Identity/ accession no.	Organism	Putative function	Score	E value	
SNP1	LjSGA_024383.1	Amino acid transporter family protein	50	9e-06	ABD32531.1	*Medicago truncatula*	Amino acid/polyamine transporter II	181	1e-53	Lys to Asp
SNP2	chr6.CM0437.400.r2.m	Unknown protein	722	0	ABE88111.1	*Medicago truncatula*	Hypothetical protein	81.6	3e-13	Tyr to STOP
SNP3	LjSGA_14121.1.1	Polynucleotidyl transferase	56	2e-07	ABN08587.1	*Medicago truncatula*	Polynucleotidyl transferase	139	6e-38	Gln to Lys
SNP4	chr3.LjT08001.160.r2.m	Integrase, catalytic core	216	7e-56	CAN83584.1	*Vitis vinifera*	Hypothetical protein	191	5e-57	Asp to Tyr
SNP5	LjSGA_024407	Small glutamine-rich tetratricopeptide repeat-containing protein 2	170	4e-42	ACI31549.1	*Glycine max*	SGT-1	142	4e-40	Glu to STOP
SNP6	chr4.LjB06H14.20.r2.m	Plasma membrane receptor-like kinase	502	1e-142	ACF70844.1	*Medicago truncatula*	Plasma membrane receptor-like kinase	987	0	Ala to Asp
SNP7	LjT27E22.50.r2.a	Phosphatidate cytidylyltransferase	355	1e-97	XP_002516647.1	*Ricinus communis*	Phytol kinase 1	362	3e-123	Phe to Leu
SNP8	chr5.CM0325.50.r2.m	Unknown protein	222	1e-57	XP_002282285.1	*Vitis vinifera*	Hypothetical protein	287	3e-96	Phe to Leu
SNP9	chr2.CM0081.690.r2.m	Homeodomain-like protein	82	3e-15	ABD32664.1	*Medicago truncatula*	Homeodomain-like protein	324	2e-106	Pro to Thr
SNP10	chr1.CM2104.10.r2.a	Electron transport family protein	377	1e-104	XP_2882593.1	*Arabidopsis lyrata*	Electron transport SCO1/SenC family protein	357	6e-121	Ser to Tyr
SNP11	chr5.CM1574.1030.r2.m	Polynucleotidyl transferase	303	4e-82	ABM55244.1	*Beta vulgaris*	Polynucleotidyl transferase	144	6e-38	Arg to Leu
SNP12	chr4.CM0229.150.r2.m	Unknown protein	94	7e-19	AAF13073.1	*Arabidopsis thaliana*	Putative retroelement pol polyprotein	49.7	3e-03	Pro to Gln

Table 4.3: Variant and alignment output of each SNP with their published report related with ABA.

SNP name	Base change	Alternate allele no.	Variant allele (read alignment)	Putative function	Report related to ABA	Reference
SNP1	C-T	3	None	Amino acid/polyamine transporter II	None	Not found
SNP2	G-A	4	None	Hypothetical protein	None	Not found
SNP3	G-A	3	None	Polynucleotidyl transferase	None	Not found
SNP4	G-A	3	None	Hypothetical protein	None	Not found
SNP5	C-T	3	None	Small glutamine-rich tetratricopeptide repeat-containing protein	Expressed in response to exogenous ABA	Clément et al. (2011)
SNP6	**G-A**	**4**	**Occur only in *Beyma***	**Plasma membrane receptor-like kinase**	**ABA insensitive mutant in Arabidopsis**	**Osakabe et al. (2005), Osakabe et al. (2010)**
SNP7	C-T	8	None	Phosphatidate cytidylyltransferase	None	Not found
SNP8	G-A	3	None	Hypothetical protein	None	Not found
SNP9	C-T	3	None	Homeodomain-like protein	Related	Himmelbach et al. (2002), Son et al. (2010)
SNP10	**C-T**	**3**	**Occur only in *Beyma***	**Electron transport family protein**	**None**	**Not found**
SNP11	G-A	3	None	Polynucleotidyl transferase	None	Not found
SNP12	C-T	3	None	Unknown protein	None	Not found

*Bold SNPs were found to be real mutation in *Beyma*.

4.4.3.2 Procedure 2

More than 220,000 SNPs/indels were called from each genome (Figure 4.5B). A total of 452 predicted SNPs occurred uniquely in *Beyma* (Table 4.4). The SNPs included eight indels and 42 SNPs of C-to-T and G-to-A base substitutions. Chromosome 1 had the highest number of SNPs/indels called, which was 50. A total of 245 SNPs/indels were found in the unmapped contigs. The least number of unique SNPs was in chromosome 6 (23 SNPs). Regardless of looking at the EMS-canonical base substitution, 50 SNPs were located in coding regions. Thirty six SNPs caused amino acid changes including three SNPs with C/G to T/A changes. None of the indel mutations were located in coding regions, which resulted in a nonsynonymous change or frameshift mutation. Out of 36 SNPs, 23 were located in known locations in the *Beyma* chromosomes. The other thirteen SNPs were in the unmapped contigs.

Each SNP was scanned manually by analysing the mapped reads and the contigous sequences of each SNP. Thus, the number of SNPs being verified could be reduced and false positive results from the SNP calling could be avoided. Two characters of the SNPs were taken into account. First, the allele balance ratio of AB was set up at a minimum value, 0.8 (80 % of mutant allele in mapped reads). Consequently, a possibility of reference base (RB) occurrence in the mapped reads was expected at 0.2 (20 % of reference allele in mapped reads) or fewer. Therefore a SNP was disregarded if RB occurred in the mapped reads.

Second, whole-genome duplication of *L. japonicus* showed that 13 % of genes assigned in six large duplicated regions between chromosomes were conserved in each pair of duplicated segments (Sato et al., 2008). The duplication affects the read mapping and causes the alignment of non-allelic homologous reads, which resulted in read mis-pairing and false positive SNPs. Therefore if mismatched bases occurred in the contiguous sequences of the same or different reads of a SNP, the SNP was then removed. This scanning narrowed down the number of SNPs to three (Table 4.5). One SNP were located in chromosome 3 and caused a change from glutamic acid to lysine in an F-box family protein. The other two SNPs were found in the unmapped regions in which leucine-to-valine change in an ethylene insensitive-like protein and serine-to-phenylalanine change in a methyltransferases superfamily protein.

Table 4.4: Number of SNPs/indels called in each chromosome and unmapped contigs (procedure 2).

Chromosome	No. of SNPs/indels predicted
1	50
2	32
3	37
4	38
5	27
6	23
Unmapped	245
Total	452

Table 4.5: List of SNPs found to be unique to *Beyma* (procedure 2).

No	Chromosome	Contigs	Locus	Putative function	Base change	AA change	
						WT	*Beyma*
1	3	CM0451	1060.r2.d_1	F-box family protein	G-A	Glu	Lys
2	Unmapped	LjSGA_055593	1_1	Ethylene insensitive-like protein	A-C	Leu	Val
3	Unmapped	LjSGA_059574	1_2	Methyltransferases superfamily protein	G-A	Ser	Phe

4.4.3.3 Procedure 3

This procedure identified 734 SNPs as background mutations, which occurred in both *Beyma* and WTS. A total of 719 SNPs were unique in *Beyma*. SnpEff identified 57 SNPs predicted to change amino acid sequence or produce nonsynonymous substitution. One SNP affected splice donor and splice acceptor sites each (Table A1). The nonsynonymous SNPs were distributed randomly in the genome. One SNP was located in an unmapped region. The highest number of nonsynonymous SNPs occurred in chromosome 3 (twelve SNPs), followed by chromosome 5 with eleven SNPs. Chromosomes 2 and 4 had ten and nine SNPs, respectively. Meanwhile, chromosomes 1 and 6 had seven SNPs each. One SNP located at chromosome 3 (chr3.CM0451.1060.r2.d) was also identified in procedure 2. This locus encodes for an F-box family protein. Ten out of 57 putative nonsynonymous SNPs occurred in genes encoding unknown proteins. Two SNPs affected the amino acid sequence of genes, which

are involved in ABA biosynthesis (*ABA DEFICIENT 2* and *ABSCISIC ACID 8'-HYDROXYLASE*). These two SNPs were located at different loci, which were analysed in Chapter 3. Furthermore, the SNPs occurred in splice site acceptor and donor of genes encoding *BROMODOMAIN-CONTAINING FACTOR 1* and *WD-40 REPEAT FAMILY PROTEIN*, respectively. The other affected genes encode for various classes of protein functions.

4.4.4 Putative causative mutation

Before successful isolation of WTS in this study, SNP verification could only be done on the *Beyma* genome. In order to confirm if the SNP was a true mutation, PCR sequencing was carried out on F2 population of an outcross between *Beyma* and Gifu. This procedure was performed on SNP6 (chr4.LjB06H14.20.r2.m) encoding a plasma membrane receptor-like kinase protein. In *A. thaliana*, the expression of a membrane-bound *RPK1* was induced by ABA and dehydration (Hong et al., 1997) and mutation in the *RPK1* gene enhanced resistance to ABA (Osakabe et al., 2005; Osakabe et al., 2010). Osakabe et al. (2005) inserted T-DNA mutation into the *A. thaliana RPK1* gene (locus: AT1G69270) which exhibited an ABA insensitive phenotype. In this study, the SNP6 locus sequence identified did not match to the same locus that was mutated by Osakabe et al. (2005) and Osakabe et al. (2010). The SNP6 locus showed high similarity to the *A. thaliana* locus, AT5G63710, with similar molecular function which is involved in serine/threonine kinase activity (TAIR). As such, the SNP locus was designated as *LjRPK1-like* gene.

The genomic DNA of *LjRPK1-like* is 4,483 bp in full length and consists of ten exons with 1,788-bp total coding sequence. The predicted protein length is 595 amino acids (Figure 4.6). The mutation occurred in the ninth exon with a G-to-A base change. Consequently, the mutation of GCT-to-GTT codons (Alanine-to-Valine) occurred in its kinase domain at the downstream region. The *LjRPK1-like* gene showed high similarity with a serine/threonine protein kinase in *M. truncatula* (Medtr4g144240) and *G. max* (Glyma05g33000) based on the Phytozome database. They possess similar protein domains that are comprised of a signal peptide at the N-terminal, four motifs of leucine-rich repeat (LRR) but *M. truncatula* has an extra LRR motif and a kinase domain at the C-terminal domain (Figure 4.7). The length of their kinase domain is slightly different; *LjRPK1-like* has the shortest length (268-aa) while *MtRPK1-like* and *GmRPK1-like* have 274-aa and 291-aa, respectively (Figure 4.6).

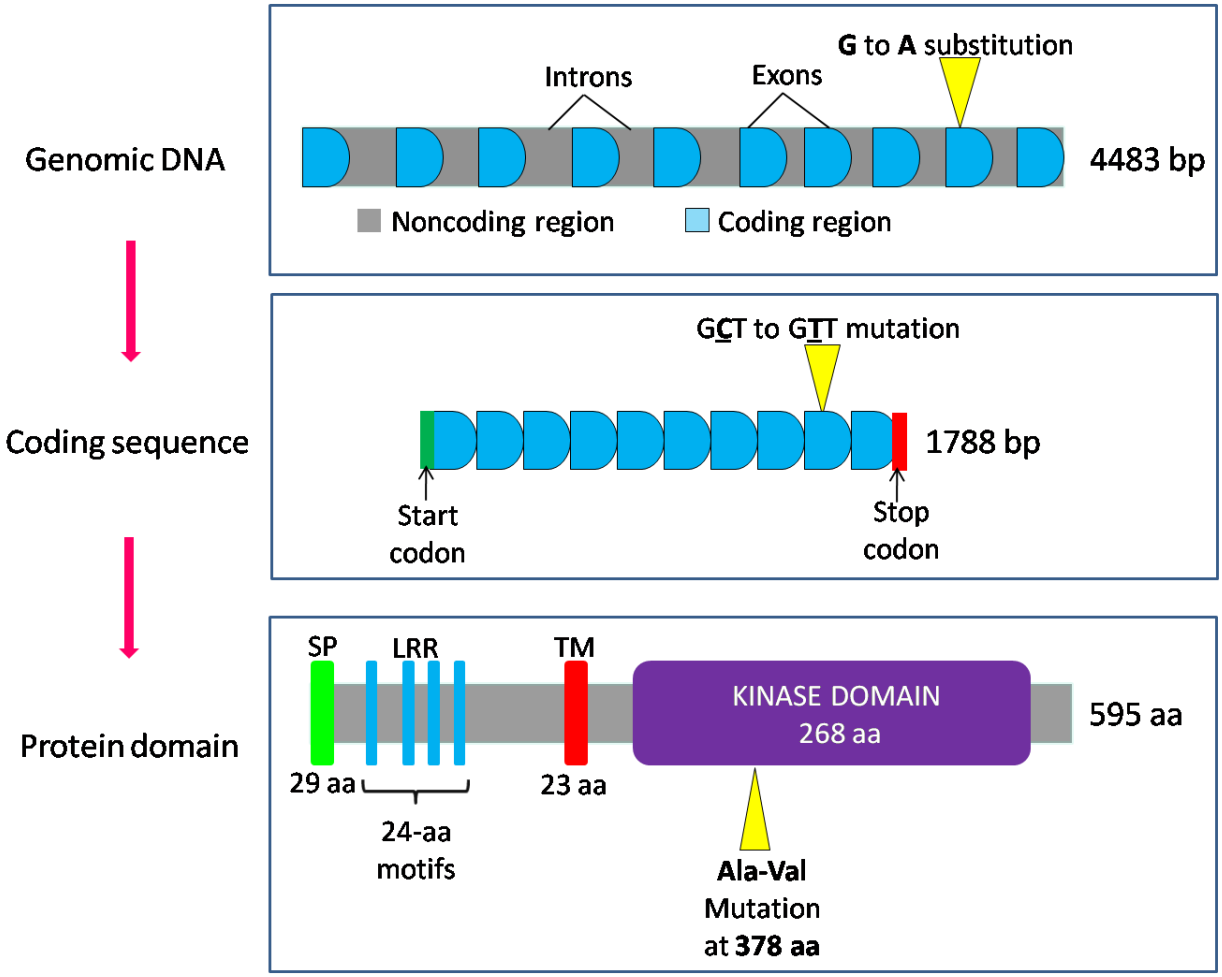

Figure 4.6: Illustration of gene structure of *LjRPK1-like*. This gene consists of 4,483 nucleotides with ten exons and the mutation occurred in the ninth exon which caused the alteration of Alanine-to-Valine amino acids in the kinase domain. SP, signal peptide; LRR, leucine-rich repeat; TM, transmembrane; aa, amino acid.

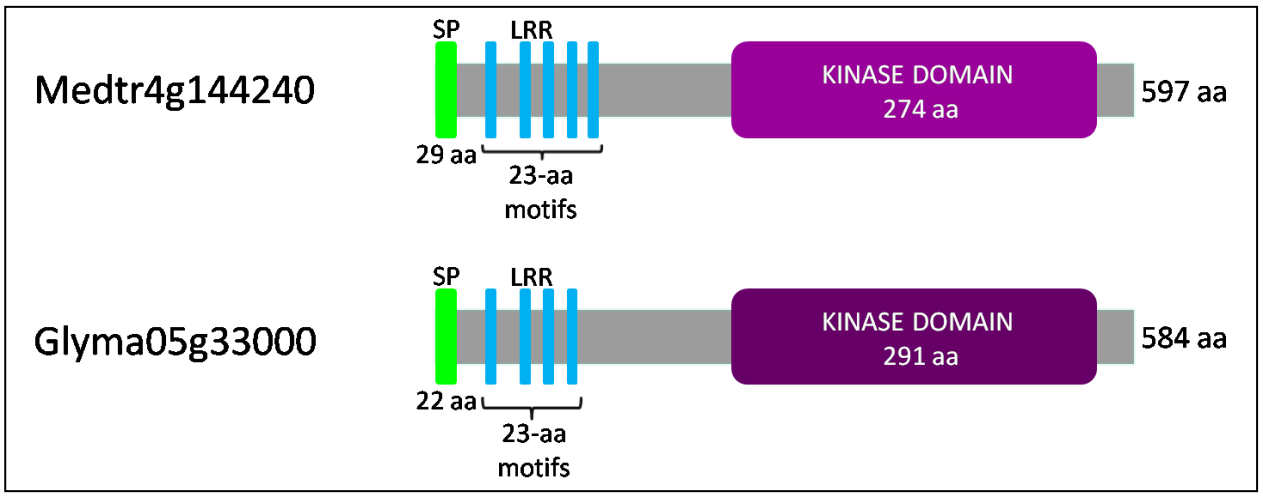

Figure 4.7: Illustration of protein domains of *RPK1-like* orthologous gene in *M. truncatula* (top) and *G. max* (bottom). SP, signal peptide; LRR, leucine-rich repeat; aa, amino acid.

4.4.5 Verification of causative mutation in F2 plants

PCR-sequencing of *LjRPK1-like* locus was carried out on F2 plants. As a preliminary test, only fifteen F2 individuals were randomly selected from the F2 population of 50 dehydrated plants. This test aimed to show if the SNP occurred in the *LjRPK1-like* locus. As comparison, WT (Gifu and MG-20), *Beyma* and F1 plants (generated from a crossing of *Beyma* and Gifu) were also subjected to the sequence analysis (Table 4.6). The variant allele (A) was observed in the *Beyma* lines as expected. Meanwhile, there was no variant allele in Gifu and MG-20 (reference allele was G) as they were unaltered genomic DNA plants (Table 4.7). Out of 15 F2 plants, the PCR-sequences of *LjRPK1-like* locus showed heterozygosity of seven plants. Four F2 plants displayed homozygous WT alleles. The other four of F2 plants showed homozygous mutant alleles. As a result, a ratio of homozygous mutant to heterozygous to homozygous WT (AA:AG:GG) was obtained as 1:1.75:1.

Table 4.6: PCR-amplified sequencing results of WT, *Beyma*, F1 and F2 plants.

Plant	Sequence name	Allele		Genotype
		Reverse	Forward	
F1	F1-1	G/A	G/A	Heterozygous
	F1-2	G/A	G/A	Heterozygous
Gifu	G1	G	G	Homozygous WT
	G2	G	G	Homozygous WT
MG-20	MG	G	G	Homozygous WT
Beyma	B3	A	A	Homozygous mutant
	B4/1	A	A	Homozygous mutant
	B4/2	A	A	Homozygous mutant
F2	0.65	G/A	G/A	Heterozygous
	16	G/A	G/A	Heterozygous
	48	G/A	G/A	Heterozygous
	128	G/A	G/A	Heterozygous
	411	G/A	G/A	Heterozygous
	433	G/A	G/A	Heterozygous
	314	G	G	Heterozygous
	32	G	G	Homozygous WT
	36	G	G	Homozygous WT
	225	G	G	Homozygous WT
	435	G	G	Homozygous WT
	115	-	A	Homozygous mutant
	213	-	A	Homozygous mutant
	223	-	A	Homozygous mutant
	318	-	A	Homozygous mutant

4.4.6 Verification of putative causative SNPs in the mutants

Five putative causative SNPs identified in procedures 1 and 2 were PCR-sequenced (Table 4.7). It showed that bases were mutated in the *Beyma* and WTS genomes, indicating they were background mutations. Verification of SNPs from procedure 3 had not been performed due to time constraint.

Table 4.7: Output from PCR sequencing of putative causative SNPs obtained from procedure 1 and 2.

No	Locus name	Base change	Base		
			Beyma	WTS	WT
1	chr4.LjB06H14.20.r2.m	G-A	A	A	G
2	Chr1.CM2104.10.r2.a	C-T	T	T	C
3	chr3.CM0451.1060.r2.d*	G-A	A	-	G
4	LjSGA_055593	A-C	C	C	A
5	LjSGA_059574	G-A	A	A	G

*Verification is in progress.

4.5 Discussion

4.5.1 Phenotyping of F2 plants

In the presence of ABA or drought stress, guard cells control the loss of water content by triggering a reduction of turgor pressure and consequently, cause stomatal closure (Sirichandra et al., 2009; Cutler et al., 2010; Ng et al., 2014). However, the seedlings of ABA insensitive *Beyma* grew normally in ABA-supplemented medium and the guard cells remained open under ABA treatment. This condition caused the dryness susceptibility observed in *Beyma* (Biswas et al., 2009). In this study, the dryness tests were applied to select F2 plants carrying the homozygous mutated gene. The correlation between the effect of ABA and drought stresses was taken into consideration. ABA interacts with many genes and cross talks with other hormones in regulating the events of plant physiology (as reviewed in Kermode 2005; Fujita et al., 2006; Rock et al., 2010). Thus, a ratio of 1:1.7 and 1:1.3 for leaf and shoot wilting showed a possibility of a pleiotropic effect of the mutated gene.

PCR-sequencing of the *LjRPK1-like* locus containing a putative causative SNP was performed on the fifteen F2 plants putatively carrying homozygous causative mutated alleles to observe segregation of the causative SNP. Our study showed that a ratio of 1:1.75:1 for homozygous mutant to heterozygous to homozygous WT alleles was obtained, indicating that the *LjRPK1-like* gene was a dominant mutation. Further verification of this gene as a causative candidate was carried out by PCR-sequencing on the WTS genome later.

However, the phenotyping test should be repeated by applying ABA treatment on the F2 population in order to obtain a reliable result and avoid a false positive. In addition, *Beyma* is a dominant mutation (Biswas et al., 2009). The selection of homozygous *Beyma* F2 plants could be biased to F2 carrying heterozygous *Beyma* alleles too. In this case, F2 carrying Gifu allele should be selected for mutation verification. This issue could have affected the selection results. Moreover, the age of the seeds and the associated low germination rate could also affect the results.

Initially, rough mapping analysis was performed by selecting two SSR markers that produced more than 20 bp amplicons (to facilitate viewing using TAE gel ecletrophoresis) from each chromosome. After running nine markers on 50 F2 plants, the markers did not show a skewed segregation to MG-20 (data do not shown). Considering the unreliable phenotyping result, the SSR mapping analysis was aborted. Therefore, new outcrossing of *Beyma* and Gifu was performed (Chapter 5) in order to obtain a new F2 population for further analysis.

4.5.2 Identification of putative causative SNPs

Although the cost of whole genome sequencing is drastically reduced with the existing high throughput technology, bigger challenges come to the analysis and interpretation of the sequencing data (Pabinger et al., 2013). Many tools have been developed to run data analysis, such as quality assessment, alignment, variant identification and visualisation (Zhang et al., 2011; Pabinger et al., 2013; Yu and Sun, 2013). Here, three procedures using different sequencing data analysis tools were performed to identify a *Beyma* causative mutated gene. The large number of SNPs in procedures 1 and 2 was caused by lenient parameters set up during the SNP calling. However, the number was reduced after screening for unique SNPs in the *Beyma* to remove background mutations and false positives. All procedures also demonstrated

variant outputs because variant tools use varied algorithmic rules in calling variant or SNPs, resulting in different number of SNPs at the same or different positions in the genome (Pabinger et al., 2013; Yu and Sun, 2013).

Reference assembly and SNP calling tools not only affected the number of called variants or SNPs, but also the identification of desired variants or causative SNPs (Pabinger et al., 2013; Yu and Sun, 2013). Our analysis showed that different putative causative SNPs were identified from different procedures except one SNP, which was located in a locus named chr3.CM0451.1060.r2.d. This SNP was listed as a putative causative SNP in procedures 2 and 3 as well as in the re-sequencing of *Beyma* and WTS in Chapter 5. Verification of this locus is still in progress. These results showed the discordance in different assemblers or variant callers, leading to the SNP chr3.CM0451.1060.r2.d not to be identified in procedure 1.

Different reference assembly tools utilised distinct algorithms and default parameters on trimming, quality scores, paired end read length and mapping (Zhang et al., 2011). Moreover, SNP calling algorithms could be distinguished in their filtering criteria such as coverage cutoffs, quality scores and read hits (Pabinger et al., 2013; Yu and Sun, 2013). For example, more than three reads was set as the coverage cutoff to call as SNP in procedure 1, but procedure 3 accepted as low as two. Procedure 2 began with calling SNPs with less than 20 % mismatches at coverage cutoff of three, prior to removing SNPs with mismatched reads at coverage of at least one. This case could also be the reason why SNP chr3.CM0451.1060.r2.d was identified only in procedures 2 and 3. In addition, SGSautoSNP disregarded the read quality score in its algorithm, unlike FreeBayes (Lorenc et al., 2012; Garisson and Marth, 2012). These algorithmic changes also gave impacts on the variant or SNP calls in different read depths, leading to a missing of a number of SNPs (Chapman, 2013). Nevertheless, SGSautoSNP produces a high confidence level of SNPs with accuracy greater than 93 % in wheat (Lorenc et al., 2012) and more than 97 % in canola (J. Batley, personal communication). With higher confidence of assembly and variant caller tools as well as output from procedure 3, it was also adopted in analysing our re-sequencing data (described in Chapter 5).

As discussed in Chapter 2, EMS mutation has biased changes of G/C-to-A/T bases (Lawley and Martin, 1975; Sega, 1984; Haughn and Somerville, 1987; Sikora et al., 2011). In addition, Greene et al. (2003) and Perry et al. (2009) previously showed more

than 97 % of G/C-to-A/T mutation in *Arabidopsis* and Gifu, respectively. This influenced the parameter set in procedure 1 to focus on the EMS canonical base substitution. Meanwhile, procedures 2 and 3 did not restrict to the EMS canonical base substitution in order to expand the identification of causative SNPs. In addition, analysis of EMS effects on the single plant genome of EMS mutagenised *L. japonicus* (Chapter 2) showed that frequencies of G/C-to-A/T changes were not too high, which were 45 % (WTS) and 34.9 % (*Beyma*). Similar results were also reported in other plants, wherein 70 % in barley (Caldwell et al., 2004), 70 % in rice (Till et al., 2007) and 60 % in tomato (Minoia et al., 2010). Thus, the identification of causative SNP should not be limited to the EMS canonical base changes in this project.

4.5.3 Background mutation

In vivo verification is a prerequisite to confirm putative SNPs in our mutants as well as to ensure they are not background mutation. In this study, our mutants were derived from the same EMS-mutagenised seed population (Figure 4.1). EMS randomly affected the *Beyma* and WTS genome sequences, in which the generation of heterozygous and homozygous mutated alleles could have occurred (Henry et al., 2014). Thus, they might carry the same or different SNPs, either heterozygous or homozygous, which randomly segregated during meiosis (Snustad and Simmons, 2003). This can create a possible scenario of mutation where; a noncausative SNP was homozygous in *Beyma*, but heterozygous in the WTS genome. Thus, the SNP could have only been identified in *Beyma*. Our SNP calling only identified homozygous mutations in order to avoid in calling false positive SNPs (Lorenc et al., 2012). As *Beyma* mutation is a dominant, the causative SNP will not appear in WTS either as heterozygous or homozygous. These parameters resulted in the identification of false causative SNPs, which were actually background mutation.

In addition, different *Beyma* and WTS plants were used for deep sequencing and *in vivo* verification. During mutagenesis, EMS creates random point mutation in the mutagenised seeds, which inherit base changes to their progenies (Greene et al*.,* 2003; Sikora et al*.,* 2011). In this study, a single plant genome of *Beyma* and WTS was deep sequenced. The causative SNP could always be identified in the *Beyma* mutants but should not occur in WTS as described earlier. A noncausative SNP could be present or absent in the different WTS plant genomes due to segregation, which doubted the reliability of SNPs in this case. Hence, the same WTS plant is required to be utilised for the

in vivo verification. However, the WTS plant sample used for the deep sequencing was unable to be traced in this study. Alternatively, both *Beyma* and WTS were re-sequenced (Chapter 5) in order to achieve the objective of this project in identifying the *Beyma* gene as well as to obtain a better quality of sequencing data.

4.5.4 Potential causative mutation

The application of different sequencing data analysis tools and parameters has resulted in variant outcomes, providing several good chances in finding a causative mutation in the ABA insensitive *Beyma* mutant. Prior to analysing the re-sequencing data of *Beyma* and WTS (Chapter 5), procedure 3 was performed on their single plant genomes. As a result, a new list of putative causative SNPs was obtained. Unfortunately, *in vivo* verification of the SNPs could not be carried out due to limited period of time and unavailability of the original WTS sample. A new minor project could be proposed to verify whether these putative causative SNPs are a real causative mutation or not. In addition, a quicker verification by demonstrating SNP mapping could also be attempted on a segregant population of an outcross between *Beyma* and Gifu (Chapter 5).

In this chapter, the number of potential causative mutations was reduced by selecting mutations or SNPs that predictably led to amino acid changes only, which reduced the candidate number. However, the searches would be extended to other SNPs, which were located in the upstream or downstream regions of the annotated MG-20 genes, if the actual *Beyma* mutation does not occur in the translated regions. A few potential causal mutations had impaired proteins, which were reported to be related with ABA, such as ABA 8'-hydroxylase proteins that partake in stomatal movement by controlling ABA catabolism in guard cells and vascular tissues (Okamoto et al., 2009). Other examples are leucine rich repeat receptor like kinase, *RPK1* (Osakabe et al., 2005; 2010), F-box containing gene, *MAX2* (Bu et al., 2013) and pentatricopeptide repeat containing gene, *SOAR1* (Mei et al., 2014) involved in ABA signaling of *A. thaliana*.

As many mutational analyses backcrossed an interest mutant to its isogenic WT combining with a DNA pooling analysis (Ashelford et al., 2011; Mokry et al., 2011; Hartwig et al., 2012), we demonstrate that an approach of single genomic comparative analysis between our mutant, *Beyma*, and its WTS produced a promising outcome. Our results also showed that a backcrossing is not prerequisite to identify a causal gene, in which save time and cost. In addition, EMS-induced mutations could be observed throughout the

genome as described in Chapter 2. Moreover, a survey of background mutation has been hitherto overlooked in genome-phenome linkages. Thus, analysing the EMS effects on the *Beyma* and WTS genomes can show not only the causal gene but also, genes associated with other traits.

4.6 Conclusion

The *Beyma* gene presumably has pleiotropic effects on *L. japonicus* based on the dehydration tests. As a result, phenotyping should be repeated by selecting homozygous Gifu phenotype to obtain a reliable result. Three different procedures of sequencing data analysis resulted in a number of potential causative mutations. Five putative causative SNPs have been shown as background mutations due to EMS mutagenesis of MG-20. Later, a total of 59 putative causative SNPs were identified and predicted to affect amino acid sequences of the *L. japonicus* genome. In future, a validation test should be carried out to determine if they are a causative mutated gene or background mutation.

Chapter 5

Re-sequencing of the *Beyma* and WTS genomes to identify an ABA insensitive *Beyma* gene

5.1 Abstract

Current sequencing technology offers better generating systems to undertake whole genome sequencing rapidly and efficiently. Development of the sequencing devices grows together with data analysis mechanisms to accomplish an objective in forward genetics and genomics studies. In this project, we had sequenced our mutants of interest to discover a causal gene in the ABA insensitive *Beyma* mutant. Re-sequencing of the *Beyma* and wild type segregant of the *Beyma* (WTS) genomes was later performed to improve the output data obtained and intensify the identification of the causal gene. The re-sequencing was carried out on pooled DNA using a different sequencing platform, Illumina MiSeq. Data analysis was performed using procedure 3. Mutation frequency of both mutants increased ~18-35 %. Unique *Beyma* mutations also rose up to 31 % of the individual sequencing output, demonstrating that pooled DNA sequencing increased the mutation frequency. There were 69 unique *Beyma* SNPs predicted as nonsynonymous alteration and will be verified in future study. Nevertheless, a mutation of C-to-T change (locus named chr3.CM0451.1060.r2.d) was found in both batches of sequencing in an F-box family gene. This gene could be the *Beyma* gene but it requires verification. In addition, a F2 population of outcross between *Beyma* and *L. japonicus* ecotype Gifu was also prepared for the segregation analysis of putative causal SNPs. Outcome of this project might provide data for sequence analysis in legumes and mutational analysis. In addition, the identification of the causal *Beyma* gene possibly identifies a novel gene involved in ABA sensitivity in legume systems.

5.2 Introduction

Rapid and cost-effective next generation sequencing (NGS) accelerates whole genome sequencing and facilitates the discovery of mutation-induced polymorphisms in plant forward genetics (Mardis, 2007; Zhang et al., 2011), contributing to the development of crop breeding program (Varshney et al., 2009, 2014; Edwards et al., 2013). Bioinformatics analysis tools have been intensively developed to provide reliable and effective systems. This technology allows *de novo* sequencing, whole genome sequencing, exome sequencing and re-sequencing of reference genomes or non-model genome without sequence information (Thudi et al., 2012; Pabinger et al., 2013; Yu and Sun, 2013). However, the identification of mutation that is responsible for a phenotypic variation could be a challenging task especially in an incomplete genome sequence as well as due to genome complexity (Nordström et al., 2013). Yet, many genes have been established as causal mutations via NGS technologies in various species such as *Arabidopsis thaliana* (Uchida et al., 2011; Austin et al., 2011; Tabata et al.,2013), *Caenorhabditis elegans* (Zuryn et al., 2010), *O. sativa* and *Arabis alpina* (Nordström et al., 2013).

Genome sequencing platforms also play essential roles in obtaining high throughput data with high accuracy. Initially, sequencing technology was firstly introduced by Frederick Sanger using dideoxy chain termination mechanism (Sanger et al., 1977). At present, a number of NGS instruments are available offered by different companies such as Illumina, Life Technologies, PacBio and Roche. These NGS platforms are varied in cost as well as accuracy and run different mechanisms including sequencing-by-synthesis (Illumina Genome Analyser/HiSeq/Miseq), pyrosequencing (Roche 454), and oligonucleotide probe ligation (Life Technologies SOLID) principles (Pareek et al., 2011; Liu et al., 2012; Thudi et al., 2012). The availability of various NGS platforms offers researchers to choose a good performance system, which is rapid, reliable and substantially low cost.

In this project, we aimed to identify a causal gene in ABA insensitive *Beyma* mutant (Biswas et al., 2009) by implementing NGS technology. We performed two batches of whole genome sequencing using different platforms, which were Illumina GA IIx and MiSeq. The first batch involved the sequencing of a single individual each of *Beyma*

mutant, wild type segregant of *Beyma* (WTS) and wild type (WT) of *Lotus japonicus* ecotype Miyakojima (MG-20) using Illumina GA IIx, producing more than 25 millions paired reads of 100 bp each (Chapters 2 and 4). A total of 57 SNPs were identified as putative causal mutations, which were predicted leading to nonsynonymous changes in various genes of MG-20 (Chapter 4).

Our first sequencing batch was done in 2010 (P. Gresshoff, personal communication) and later, a benchtop sequencer MiSeq was launched in 2011 (Liu et al., 2012). Although GA IIx and MiSeq are able to produce a significant yield of bases greater than quality of 30, MiSeq requires a shorter time (~27 hours) to run its workflow including cluster generation (Quail et al., 2012; Thudi et al., 2012). The utilisation of MiSeq opted to generate up to 250 bp paired end reads (PE; http://systems.illumina.com/systems/miseq/performance_specifications.html) as compared to GA IIx that only generates up to 150 kb PE (Liu et al., 2012). In addition, longer read length was reported to reduce the complexity of sequence assembly (Koren et al., 2013). The improvement of current NGS technology also provides a better system to obtain a good quality of sequencing output. Taken altogether, we decided to re-sequence pooled DNA of *Beyma* and WTS to intensify the identification of the causal mutated gene. This chapter presents the outcomes of the re-sequencing as well as the preparation of samples required for further analysis of the potential causal gene.

5.3 Materials and methods

5.3.1 Outcrossing between *Beyma* and *Gifu*

Gifu pollens were transferred to *Beyma* stigmas as described in Jiang and Gresshoff (1997). A total of 17 *Beyma* flowers were pollinated with *Gifu* pollens. Seeds produced from the fertilisation were germinated and allowed to grow for two months before DNA extraction. Genomic DNA of each crossed plants was amplified using two SSR markers (chr1TM0231 and chr4TM0266; http://www.kazusa.or.jp/lotus/clonelist.html) to validate the crossing.

5.3.2 Isolation of new WTS plants

Two ABA treatments were performed on two EMS-mutagenised M3 populations (2538-1 and 2538-2) to isolate seeds with WT phenotype (sensitive to ABA) which were then called WTS. First treatment (ABA on germination), EMS-mutagenised seeds were germinated on filter paper wetted with and without (control) 100 µM ABA for 5 days (Biswas et al., 2009). Non-germinated seeds were transferred to wetted filter paper without ABA to allow them germinate for 3 days before transferring onto half strength B5 medium for 7 days. Second treatment (ABA-root assay), ABA-free germinated seedlings (with the length of ±3 mm radical or/and 5-10 mm radical roots) were transferred onto half strength B5 medium supplemented with and without (control) 50 µM ABA (Biswas et al., 2009) for 7 days. Germination rate of seeds and root length of seedlings were scored after 5 days of germination and 7 days of growth, respectively. As controls, MG-20 and *Beyma* lines were also treated in parallel to screening mutagenised populations.

5.3.3 Genomic DNA extraction

Genomic DNA was isolated from plant tissues using CTAB method and subjected to RNase treatment as described in Chapter 2. In this Chapter, three individual plants were used for DNA extraction and genomic sequencing.

5.3.4 Re-sequencing of the *Beyma* and WTS genomes

Whole genome paired-end, 250 bp, short-sequence reads (>10x coverage) for *Beyma* and WTS were generated using the Illumina MiSeq according to the manufacturer's instructions. These two datasets and paired-end reads of MG-20 (from Chapter 2 and 4) were then mapped to the MG-20 genome (www.kazusa.or.jp/lotus/) using program SOAP2 v2.21 (Li et al., 2009).

5.3.5 Identification of putative causative SNPs in the re-sequenced *Beyma* genome

SNPs were called using SGSautoSNP 2.001 (Lorenc et al., 2012) wherein re-sequenced *Beyma*, re-sequenced WTS and WT were referred as different cultivars. In order to avoid false positive output, only homozygous SNPs were selected for further analysis. SNPs were categorised using SnpEff 3.0j (Cingolani et al., 2012) according to their effect on *L. japonicus* MG-20 annotated genes (Sato et al., 2008). These protocols were performed by Ms. Jenny Lee (ACPFG). SNPs occurring uniquely in *Beyma* were subtracted from the list and analysed for nonsynonymous and synonymous changes. Nonsynonymous SNPs were down listed for further verification.

5.4 Results

5.4.1 Identification of WTS plants

Germination of MG-20 seeds was almost completely inhibited by 100 µM ABA in this study (Figure 5.1). *Beyma* lines were represented by line B3, B4 and B5. B3 seeds selected from M5 generation (B3M5) showed poor germination in the presence of ABA due to contamination. Therefore, the germination rate of B3M5 was ignored. The germination rate of B4 line from different generations (B4M4, B4M5 and B4M6) was similar in presence and absence of ABA, indicating that B4 is a homozygous mutant line. B5M4 also showed reduced sensitivity to ABA during germination (Figure 3). Two groups of screening population (2538-1 and 2538-2) germinated more than 50 % without ABA but behaved differently in the presence of ABA. Fifty five percent of 2538-1 seeds germinated without ABA and only 7 % germinated with ABA present. Meanwhile, 68 % and 33 % of 2538-2 seeds germinated in ABA absence and presence, respectively. In order to rescue WTS plants, non-germinated seeds were transferred from ABA to ABA-free wetted filter paper before growing onto B5 medium. Two seedlings of each screening population were selected from this test after rescuing for further analysis.

There were two batches in the ABA-root assay; seedlings with 3-mm radical roots (Batch 1) and seedlings with 5-10 mm radical roots (Batch 2). The average of root lengths after 7 days of growth on B5 medium supplemented without and with 50 µM ABA for both batches was plotted in a graph (Figure 5.2). MG-20 root growth was impaired by exogenous ABA, similarly as reported by Suzuki et al. (2004) on *L. japonicus* and *Trifolium repense*. Unlike MG-20, *Beyma* roots were not badly affected by the presence of ABA. MG-20 roots grew to an average of 6-mm and 10-mm lengths for Batch 1 and Batch 2, respectively, which were used as maximum lengths to select plants with WT phenotype in the screening population. The selected seedlings were transferred onto ABA-free B5 medium to allow their roots to grow longer before transferring to vermiculite soil. Three seedlings (2538-2) developed well and were transferred to vermiculite for further analysis. WTS plants isolated from these two treatment were subjected to DNA extraction followed by PCR sequencing of candidate SNPs. This step was performed to determine whether the SNPs were background mutations or not.

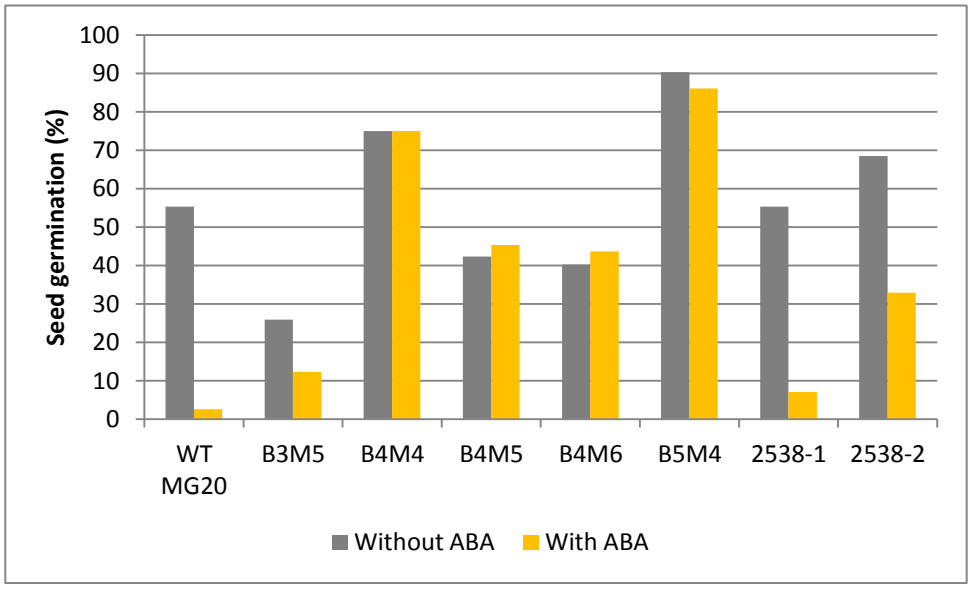

Figure 5.1: Germination rate of *L. japonicus* seeds without and with 100 µM ABA. B represents *Beyma*. M represents generation phase of mutated line.

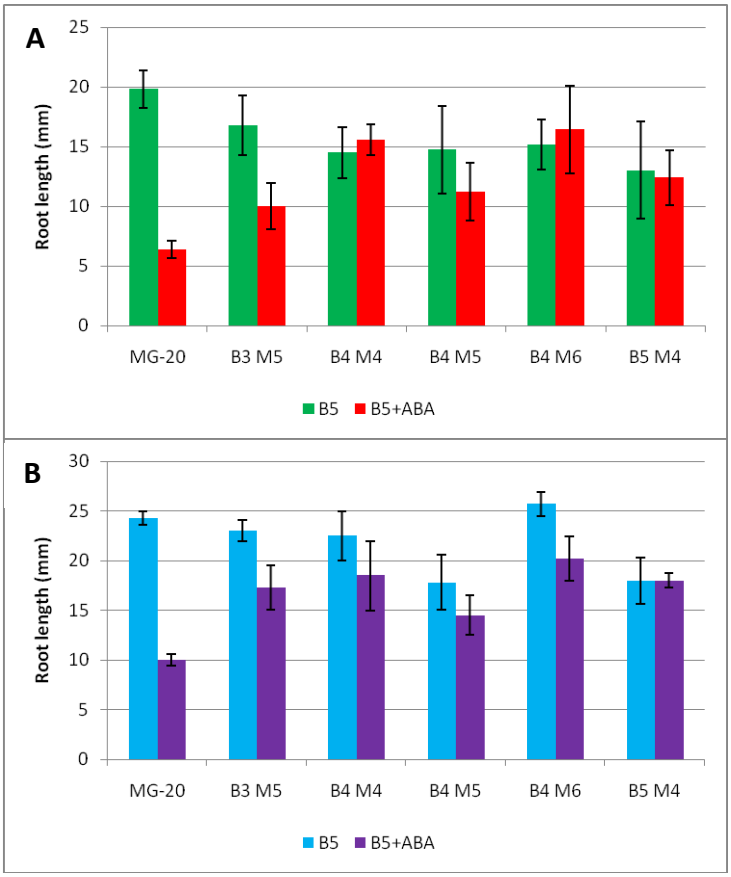

Figure 5.2: Average of root lengths grown on B5 medium supplemented with and without 50 µM ABA. Root length was measured after transferring seedlings with 3-mm radical (A) and 5-10 mm radical (B) roots to the B5 medium.

5.4.2 Verification of outcrossing between *Beyma* and *Gifu*

Out of 17 cross-pollinations, six flowers were successfully pollinated and produced pods. Four pods had more than two seeds, whilst the other two pods had only one or two seeds. One or two seeds from each pod (total of ten F1 seeds) were germinated and grown in order to confirm the outcrossing and generate F2 population. Using the TM0231 marker, five F1 plants showed two amplified bands indicating the presence of Gifu and MG-20 DNA sequences. Meanwhile, eight F1 plants showed two amplified bands using the TM0266 marker. After 3 months of growth, all these eight F1 plants produced purple mature stems due to anthocyanin deposition (Kawaguchi et al., 2001), showing that they inherited Gifu genes. Therefore, it proved that they were successfully crossed. These eight F1 plants were self-fertilised to generate F2 progenies for rough SSR profiling and SNP mapping.

5.4.3 Sequencing and read mapping output

Re-sequencing of the WTS and *Beyma* genomes produced a total of 8,169,815 and 7,446,919 paired reads, respectively (Table 5.1). Mapping procedure resulted in the mapping of 22.1 % and 20.6 % paired reads for WTS and *Beyma*, respectively, in which more 12X genome coverage. WT had the same output of paired raw reads and mapped reads as in the previous sequencing and read mapping procedures (Chapter 2).

Table 5.1: Output from read mapping of paired reads.

Genome	Paired raw reads	Read pairs mapped	% of read pairs mapped	Genome coverage*
MG-20 WT	32,965,291	9,285,440	27.27	29.88
WTS	8,169,815	1,804,550	22.1	14.52
Beyma	7,446,919	1,533,748	20.6	12.34

*Based on mapped reads

5.4.4 Frequency of mutation

The re-sequencing of *Beyma* and WTS demonstrated a slightly different mutation spectrum in their genomes. A total of 1,703 and 2,017 homozygous SNPs were identified in the WTS and *Beyma* genomes, respectively (Table 5.2). These changes resulted in a mutation rate of one mutation in every 177 kb (WTS) and 149 kb (*Beyma*). Similar to the previous sequencing batch (Chapter 2), chromosome 1 had the highest number of

mutation in both genomes, which were 401 SNPs (WTS) and 475 SNPs (*Beyma*). Meanwhile, chromosome 6 contained the least number of SNPs as shown in Table 5.2. The unmapped region and other chromosomes had a number of SNPs ranging from 200 to 330 SNPs in both mutants. Transition and transversion mutations present in both mutants were also analysed (Table 5.3). Both WTS and *Beyma* had 53.2 % and 45.1 % of G/C-to-A/T mutations, respectively, which were the highest percentage of base changes. The frequency was followed by A/T-to-G/C changes, which were 18.4 % (WTS) and 17.6 % (*Beyma*). The least frequent mutation was C/G-to-G/C changes for both WTS (2.5 %) and *Beyma* (2.4 %). This chapter focuses mainly in the identification of the ABA insensitive causal gene, thus, distribution of mutations across the whole genome strand was not analysed as in Chapter 2.

Table 5.2: Frequency of mutation and change rate occurred in each chromosome and unmapped regions of WTS and *Beyma*.

Chromosome	Length (bp)	Base changes (SNPs)		Change rate	
		WTS	*Beyma*	WTS	*Beyma*
1	66,776,104	401	475	166,524	140,581
2	44,510,304	234	305	190,215	145,935
3	48,258,781	275	304	175,486	158,746
4	43,347,107	233	326	186,039	132,967
5	37,320,184	213	219	175,212	170,412
6	28,216,978	138	166	204,471	169,982
Unmapped	32,912,371	209	222	157,475	148,254
Total	**301,341,829**	**1,703**	**2,017**	**176,948**	**149,401**

5.4.5 Unique mutations in *Beyma*

Genomic comparative analysis between *Beyma* and WTS resulted in the identification of 998 background mutations occurring in both mutants, and 940 unique mutations in *Beyma* (Table 5.4). In this subchapter, we focused on unique mutations in the *Beyma* to identify an ABA insensitive mutated gene. Data analysis showed that chromosomes 1 and 6 had the highest and lowest totals of unique SNPs, which were 198 and 67, respectively. A total of 79 SNPs were located in unmapped regions of the *Beyma*

genome. Meanwhile, chromosomes 2, 3 and 4 contained 174, 164 and 158 SNPs, which were identified only in the *Beyma*. In addition, effect of the *Beyma* mutation was also predicted in the annotated genes of the MG-20 genome (Figure 5.3). Unique mutations of *Beyma* led to a prediction of 30.9 % change effects occurring in downstream and upstream regions, each. A high percentage of mutation was also predicted occurring in intergenic regions, namely 23.4 %. In addition, only 5.6 % of mutations caused nonsynonymous effects and 0.5 % of stop-gained effects were predicted and listed (Table 5.5) for further analysis.

Table 5.3: Percentages of transition and transversion mutations in the WTS and *Beyma* genomes. High frequency of G/C-to-A/T was identified as expected.

Mutation		Changes (%)	
		WTS	Beyma
Transition	G/C-to-A/T	53.2	45.1
	A/T-to-G/C	18.4	17.6
Transversion	A/C-to-C/A	10.0	14.5
	G/T-to-T/G	10.2	13.9
	A/T-to- T/A	5.6	6.5
	C/G-to-G/C	2.5	2.4

Table 5.4: Total of SNPs identified as unique SNPs in each chromosome and unmapped region of the *Beyma* genome.

Chromosome	Changes (SNPs)
1	198
2	174
3	164
4	158
5	100
6	67
Unmapped	79
Total	940

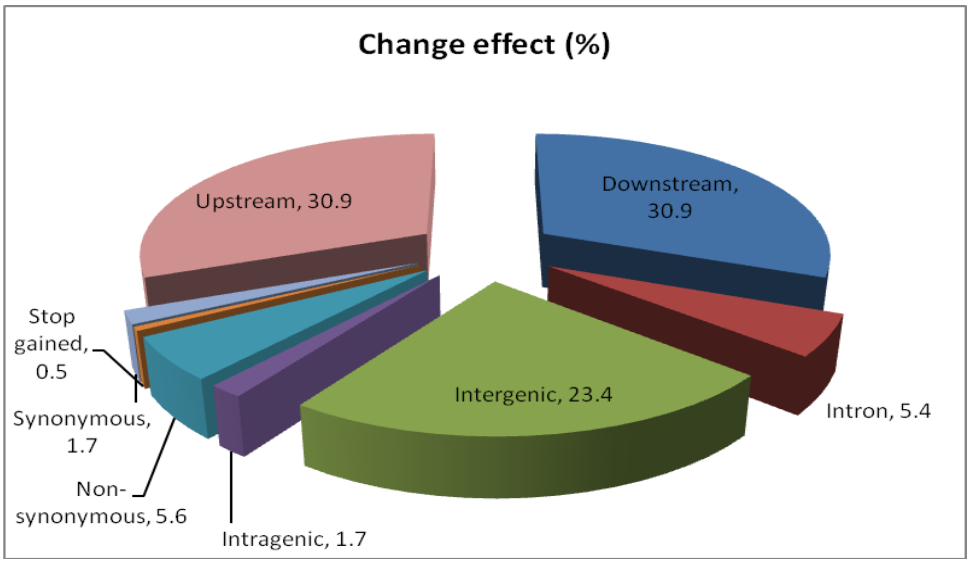

Figure 5.3: Effect of unique mutations on codon sequences in the *Beyma* genome. SNPs were observed highly located in downstream and upstream parts of the annotated genes. Only small percentage of nonsynonymous changes was predicted.

5.4.6 Putative causative mutation

Out of 940 SNPs that were found only in *Beyma*, a total of 81 SNPs were predicted leading to amino acid changes including seven stop gained mutations and 32 mutations of EMS canonical base substitutions (Table A2). Chromosomes 3 and 4 contained eighteen and twenty putative causal SNPs, respectively. Chromosomes 1 and 2 had fifteen and thirteen putative causal SNPs, respectively. Meanwhile, seven and eight SNPs were located in chromosomes 5 and 6, respectively. These putative causal SNPs located in loci that encode various annotated genes in the MG-20 genome, in which twelve of them were unknown protein. One of them was the same SNP identified as a putative causative SNP in Chapter 4. This SNP was located in chromosome 3 (locus name: chr3.CM0451.1060.r2.d), in which a C-to-T change occurred at codon number of 240. This mutation led to a change of glutamic acid to lysine in an F-box family protein. Unlike putative causative SNPs identified in Chapter 4, none of the putative causal SNPs was located in any of candidate genes as selected in Chapter 3. In addition, most of the mutations had domain or functional site of proteins such as pentatricopeptide repeat superfamily, F-box family, alpha/beta-Hydrolases superfamily, methyltransferases superfamily and kinase protein. Furthermore, stop gained mutations were located in F-box family protein, pentatricopeptide repeat superfamily protein, RHO protein GDP dissociation inhibitor, SAP domain-containing protein and unknown protein.

5.5 Discussion

5.5.1 Sample/ validating population

This chapter is not only presenting the identification of putative causative SNPs in the re-sequenced genomes, but also demonstrating the preparation of sample population for the re-sequencing and validation of putative causative mutations. Prior to the re-sequencing of *Beyma* and WTS genomes, WTS plants were screened and isolated from the original mutagenised population of MG-20 seeds. As a result, a number of plants were identified as WTS individuals with ABA sensitive phenotype. In this study, the germination ratio of screening population between ABA sensitive to insensitive showed that segregation of these populations did not follow basic Mendelian rules (Snustad and Simmons, 2003). Occurrence of background mutation might have distorted the segregation and hence affected the ratio. Since the objective of the ABA sensitivity screening treatment was to isolate WTS individuals, we omitted the germination ratio and proceeded with the selection of seeds that did not germinate on ABA as WTS genotype carriers for the re-sequencing and/or validation.

Meanwhile, the outcross products of *Beyma* to Gifu would be utilised for rough mapping using available SSR markers of the MG-20 genome or to observe segregation of the putative causal gene. Gifu was the most suitable partner due to high polymorphism between the MG-20 and Gifu genomes (>4 %; Kawaguchi et al., 2001). Since *Beyma* was derived from *L. japonicus* ecotype MG-20 (Biswas et al., 2009), crossing *Beyma* with Gifu will allow detection of mutated gene segregation in the *L. japonicus* and therefore, assist the verification of the mutated gene.

5.5.2 Re-sequencing and low genome coverage of *Beyma* and WTS

The advent of NGS technology facilitates genomics and genetics studies in many ways. This technology develops rapidly together with its analytical tools, which contribute to the improvement of experimental design that can be applied (Thudi et al., 2012; Varshney et al., 2014). Here, our attempt of re-sequencing the *Beyma* and WTS genomes was accomplished using one of the latest NGS instruments, MiSeq, on pooled individuals. We did not re-sequence the MG-20 genome because it was only used as a comparison to avoid natural variations and false positives. Thus, the MG-20 sequencing data of the first batch was re-used in this chapter. Instead of using a single genome, multiple individuals of

the *Beyma* and WTS genomes were pooled for the re-sequencing. We implied the assumption of bulked DNA segregant approach, in which a pool of mutant DNA will increase the frequency of causal mutation (Hartwig et al., 2012).

The number of mapped reads reflected the genome coverage, also called the depth of coverage (Sims et al., 2014). Our read mapping or assembly to the reference resulted in a medium depth of coverage (>12X), which was lower than the previous batch (Chapters 2 and 4). These outputs were significantly low as compared to sequencing output of other organisms, such as *A. thaliana* (Austin et al., 2011) and *Leptosphaeria maculans* (Zanders et al., 2013). The removal of multiple-aligned reads does not only increase the SNP accuracy and avoid false positives (Lorenc et al., 2012; Shiwa et al., 2012), but also decreases the number of mapped reads and the average of genome coverage (Sims et al., 2014). In addition, the assembled length of MG-20 pseudomolecules covers only 67 % of estimated genome size (Sato et al., 2008), which also affects the read mapping output.

In this study, we aimed to identify an ABA insensitive causal gene using a comparative SNP analysis of our mutagenised genomes. Our variant analysis (SGSautoSNP) called for the SNPs with ≥2 fold-coverage, leading to the identification of 32 putative causal SNPs with fold-coverage of ranging from 4-32 (Table A2). Nevertheless, SGSautoSNP produces a high confidence level of SNPs with accuracy greater than 93 % and 97 % in wheat (Lorenc et al., 2012) and canola (J. Batley, personal communication), respectively. In addition, Tabata et al. (2013) showed the identification of a causal gene in high boron requiring mutant of *A. thaliana* using low genome coverage sequencing. Thus, the low coverage of our mapped genomes would not likely affect our objective in searching of the *Beyma* gene.

5.5.3 Mutation spectrum of *Beyma* and WTS

A total of 67 % of the whole genome sequence of MG-20 was constructed in 2008 using clone-by-clone and shotgun sequencing (Sato et al., 2008), which was used for the mapping in this project. Currently, the latest version of the MG-20 genome was successfully determined using NGS technology, covering ~87 % of the total genome length (Sato and Andersen, 2014). However, the genome sequence has not been released yet. Nevertheless, genomic variation of our mutants can be determined using the available genomic sequence of MG-20. Similar to the first batch of sequencing, mutation spectrum of the re-sequenced *Beyma* and WTS genomes was analysed as compared to WT MG-20

using procedure 3 (Chapter 4). The re-sequencing of pooled mutant DNA showed that the presence of mutations was more frequent in both mutagenised genomes (~18-35 % increase), consequently resulting in higher rate of base changes. In addition, unique *Beyma* mutations were also increased ~31 % of the individual sequencing output.

These data indicated that the pooled DNA sequencing increased the frequency of mutations that are being identified, demonstrating the agreement of the assumption of bulked DNA analysis in increasing causal mutation frequency (Hartwig et al., 2012). In plant research, there are numerous NGS data obtained from pooled DNA genomes to identify genomic variation in genetic diversity or mutagenesis studies (Hartwig et al., 2012; Zhu et al., 2012b), in which the sequencing cost can be minimised. However, an individual DNA sequencing has been implied in human genome research to identify sequences that may be linked to disease or medical response prediction (Wheeler et al., 2008; Koboldt et al., 2009). Thus, the undertaking of both individual and pooled DNA sequencing provides informative output for future sequencing works in legumes.

Distribution of transition and transversion mutations did not greatly alter. Percentage of EMS canonical base substitution, G/C-to-A/T changes, was the highest as identified in the individual sequencing. Although ~97 % and 99 % of G/C-to-A/T changes were found in mutagenised Gifu and *A. thaliana*, respectively (Greene et al., 2003; Perry et al., 2009), a lower rate was obtained in other plants (Caldwell et al., 2004; Till et al., 2007; Minoia et al., 2010). This outcome has been discussed in Chapter 2, in which other base changes should not be ignored in this analysis.

5.5.4 Potential *Beyma* gene

In order to avoid too much redundancy in data analysis, this chapter presents mainly on unique mutations in the *Beyma* genome as compared to WTS and WT. In addition, the re-sequencing of our mutants was aimed to obtain a better quality data and intensify the identification of a causal gene. A large number of unique *Beyma* mutations were specifically narrowed down to mutations that were predicted leading to nonsynonymous alteration. The unique mutations occurred randomly across the genome. Quick verification by PCR sequencing will be performed in future to subtract real mutations in *Beyma*, which were absent in WT and WTS.

Nevertheless, a few putative unique mutations had a good potential as a causal *Beyma* gene. Nonsynonymous change of F-box family protein was identified occurred in the same locus in both individual and pooled DNA sequencing batches. Two other SNPs were also predicted in the same family protein. Previously, F-box containing domain genes have been reported to be linked with ABA signalling in other plants, mainly in *A. thaliana*. Similar to *Beyma*, mutants of TUBBY-like protein gene and EID1-like protein 3 showed a reduced ABA sensitivity in seed germination and early seedling development (Lai et al., 2004; Koops et al., 2011). On the other hands, a null mutation of *DROUGHT TOLERANCE REPRESSOR* encoding an F-box protein increased in drought tolerance due to ABA hypersensitivity during stomatal closing (Zhang et al., 2008; Zhang and Xue, 2009), which was contradict to *Beyma* phenotype. Interestingly, *more axillary growth 2* mutant is strongly hypersensitive to drought stress like *Beyma*, and yet, hypersensitive to ABA in seed germination and seedling development (Bu et al., 2013). This indicates the role of F-box family protein in ABA signaling could be negative or positive regulatory, demonstrating the possibility of F-box containing domain gene as a putative causal *Beyma* candidate. Other proteins, which were also nonsynonymously mutated in more than one locus, have been linked to ABA signaling. For examples, pentatricopeptide repeat family protein in *A. thaliana* (Liu et al., 2010; Mei et al., 2014) and rice (Tan et al., 2014). Yet, verification of these putative causal mutations is crucial to test for direct causality in the *Beyma* genome.

5.6 Conclusion

Re-sequencing of pooled DNA of *Beyma* and WTS showed similar mutation spectrum of transition and transversion as determined previously in the individual sequencing. Mutation frequency of both mutants and the number of unique *Beyma* mutations also increased. A total of 69 putative causal mutations were identified in our ABA insensitive *Beyma* mutant. They need to be validated as putative causal SNPs identified in the previous sequencing output.

Chapter 6
General discussion and future direction

6.1 General discussion

The advent of NGS technology accelerates the development of forward genetics and enhances the improvement of genomics studies in many species. This PhD thesis presents the application of NGS technologies in searching a causal gene in our ABA insensitive *Beyma* mutant of *L. japonicus* ecotype Miyakojima (MG-20; Biswas et al., 2009). Most of the genome sequence of MG-20 has been obtained since 2000 by the Kazusa DNA Research Institute in Japan using clone-by-clone (TAC clones) and shotgun sequencing (Nakamura et al., 2002; Kaneko et al., 2003; Asamizu et al., 2003; Kato et al., 2003; Sato et al., 2008). Currently, the latest update of this genome project has improved the length of assembled sequence, which was accomplished by modern NGS technology (Sato and Andersen, 2014). The availability of this sequence offers a good platform and resources in undertaking further genomics and mutational analysis in the model legume *L. japonicus* by adopting NGS tools. Since the latest version 3.0 has not been yet released, this project used the second version of the genome sequence (Sato et al., 2008), as a reference.

The EMS-induced *Beyma* mutant originated from a heterozygous dominant mutation (Biswas et al., 2009), which allowed the isolation of a WTS of the *Beyma* mutant containing homozygous mutant alleles. Genomic comparative analysis between these mutants could result in the identification of a causal gene in the *Beyma* genome. To date, a forward genetics study adopted a backcross of the mutant to its isogenic parent combined with a bulked DNA analysis to identify a gene that is linked to a phenotype. Here, this thesis demonstrates how analysis of single or pooled genome data sets of the mutant of interest could identify the putative causal gene by comparing genomic sequences without prior backcrossing. This approach could reduce cost and time consumed. The MG-20 genome was also re-sequenced to remove natural variation between this and reference genome.

EMS is known to be biased to G/C-to-A/T changes, which have been reported in many species, such as *Arabidopsis thaliana* (Greene et al., 2003; Till et al*.,* 2011), *O. sativa* (Till et al. 2011), *L. japonicus* (Perry et al., 2009), *Caenorhabditis elegans* (Flibotte et al., 2010; Thompson et al., 2013), *Solanum lycopersicum* (Minoia et al., 2010) and *Saccharomyces cerevisiae* (Shiwa et al., 2012) at different frequencies. The actual effects of EMS in our mutants were discovered in Chapter 2, which aimed (a) to identify SNPs in *Beyma* and WTS (as compared to re-sequenced MG-20) and (b) to show the EMS effects in the mutagenised individual MG-20 genomes. This chapter has been published in the journal G3. Although the frequency of G/C-to-A/T changes was the highest as compared to other base changes in both mutants, the percentages were relatively lower than those previously identified in *L. japonicus* by Perry et al. (2009). This result showed that the identification of the causal *Beyma* gene should not be restricted to EMS canonical base mutations. The causal gene could be impaired due to different type of base alterations, which could also give significant impact to *Beyma* phenotypes. Comparable mutation spectra between *Beyma* and WTS indicated that the sequencing of individual genomes has generated substantial output for mutational analyses and identified the presence of actual SNP loads without bias to mutations that are being discriminated from EMS collateral damages due to backcrossing.

ABA roles cover a wide range of plant systems including seed germination and responses to environmental cues, especially drought. Numerous genes have been reported to be directly or indirectly involved in the process of ABA actions from catabolism to signaling (Ng et al., 2014). Thus, a candidate gene approach was also adopted to identify the presence of mutation in orthologous sequences of selected candidates in WTS and *Beyma,* which then determined whether the candidates were putatively a causal gene in *Beyma* (Chapter 3).

EMS has impaired a small number of candidate sequences at different regions including exons and introns, either in *Beyma* or WTS or both. Unique *Beyma* mutations only occurred at a downstream part of annotated sequence of candidate loci. This approach had eliminated the candidates as the causal gene and showed the distribution of EMS effects on “ABA family genes” as being quite low. The selected candidates were mainly identified in *A. thaliana*, which has a different system in ABA-root development as compared to legumes (Liang and Harris, 2005). In addition, ABA functions are ambiguous in plant systems, suggesting that the causal *Beyma* gene could be a gene that acts

contrary to *A. thaliana* or non-legume genes or has not been characterised in legumes. Besides, preliminary analysis on *Beyma* showed that ABA inhibition is local and not involved directly in systematic autoregulation of nodulation (Biswas et al., 2009). Thus, genes that are involved in ABA-inhibition nodulation should also be selected as candidates in this chapter.

A crucial stage of this project was the sequencing and subsequent data analysis. This project involved two batches of sequencing, which utilised individual or pooled genomes of *Beyma* and WTS (Chapters 4 and 5). This effort was aimed to obtain better quality data and intensify the identification of the causal *Beyma* gene. The MG-20 genome was re-sequenced once, because it was only used for comparison in the identification of SNPs in the mutants to avoid natural variation. Three procedures of read mapping and SNP calling were attempted (Chapter 4). The first two procedures have not been previously performed in our lab. Dr Kazakoff, a former PhD student in our laboratory, carried out them as a trial-and-error procedure that successfully produced five potential causal candidates. However, they were validated as background mutations except one candidate that has not been validated in WTS. Nevertheless, these steps offered good practices in sequencing data analyses and how to deal with bottlenecks obtained in the analysing processes. During the validation, the challenge came in designing primers because the causal sequences encode for large protein family members such as F-box family protein and kinase family protein. Thus, primers need to be designed specifically to the sequences.

The third procedure (SGSautoSNP) was previously employed in sequencing analysis of the wheat genome (Lorenc et al., 2012), which was established by our collaborators in the ACPFG. They agreed to run the programs SGSautoSNP and SnpEff that work to predict mutations in annotated genes (Cingolani et al., 2012). With well-analysed data obtained from SGSautoSNP and SnpEff, the third procedure was also attempted on the output of pooled genome sequencing. This allowed relative comparison between potential causal mutations identified from both sequencing batches. Separated into two chapters, sequencing of individual and pooled genomes was described in Chapters 4 and 5, respectively. As discussed earlier, identified SNPs of individual genomes were examined thoroughly and described in Chapter 2 to address the EMS effects on a single genome of the mutagenised MG-20. On the other hand, we

emphasised the searching of the causal gene based on data from pooled sequencing of *Beyma* and WTS in Chapter 5.

Low genome coverage is not a desirable output of read mapping but it does not hinder the subsequent analyses in genomics study (Tabata et al., 2013; Sims et al., 2014). Individual sequencing generated a higher number of 100 bp reads, which consequently produced relatively high genome coverage. On the other hand, a lower number of 250 bp paired reads was retrieved from pooled sequencing, resulting in medium genome coverage. However, our coverage degrees sufficiently permitted to run the calling of mutations and SNP analysis with high fidelity (Cingolani et al., 2012; Lorenc et al., 2012). The sequencing output might also have affected by different length of paired reads and device systems utilised in this study.

Nevertheless, the same pattern of mutation spectra was observed in both *Beyma* and WTS in individual and pooled sequencing, indicating that EMS generated a comparable proportion of transition and transversion mutations during mutagenesis within the same population. In addition, DNA pooling of bulked segregants was known to increase frequency of mutations (Hartwig et al., 2012). Here, DNA was pooled from the same mutant individuals without prior backcrossing. An increase of mutation frequency was also observed, which resulted in the identification of higher change rates in our mutants. This indicates that DNA pooling is a better way to increase the frequency of causal mutation, facilitating the discovery of a causal mutated gene. Meanwhile, an individual sequencing is not only useful for the gene discovery; it also identifies the actual effect of mutagen in a genome.

Numerous genes have been established as causative mutations in induced mutants using NGS technologies and characterised in many species, especially in model plants like *A. thaliana* and *O. sativa* (Lamesch et al., 2011; Kawahara et al., 2013). To our knowledge, the application of NGS technologies in identifying causal mutated genes in the model plant *L. japonicus* and other legumes is still limited. At present, many research groups are working on the development of legume genome sequence to improve and encourage the use of NGS tools for forward genetics and genomics studies. Taking a risk of getting no results, we expended this project with many attempts and plans using the NGS approach. At this stage, we could not clarify which gene is the causal gene in our ABA insensitive *Beyma*. However, this project demonstrated that the NGS technology is

not impossible to be implemented in plants, which have incomplete gene annotation or genome sequence like *L. japonicus*. Since a unique *Beyma* mutation in an F-box family protein was appeared three times at the same position in different analyses in this project, it presumably has a good potential as the causal mutation.

On top of that, selection of samples or tissues was a minor part of this project but it needs to be carried out carefully. Homozygous *Beyma* plants were already available in our stock, which facilitated the process of its isolation. A process of isolating WTS plants was hindered by the condition of the original mutagenised population, which were old and have low rate of germination. WTS plants were managed to be isolated after a few attempts before continued with pooled sequencing. Crossing of *Beyma* and Gifu produced an F2 population that could be used for future analyses.

In summary, the identification of causal gene in ABA insensitive *Beyma* mutant is almost reached. Validation of the potential causal mutations will show the *Beyma* gene, which offers understanding and knowledge in ABA effects in legume growth and development as well as response to environmental stresses. It also provides clues in ABA-inhibition of nodulation in *L. japonicus* and other legumes. Output of this project may also contribute to EMS mutational analysis and opportunities to undertake reverse genetics study based on collateral damages identified. Not only focusing on NGS methods, knowledge on phenotypes of different *L. japonicus* ecotypes was gained and experience on handling physiological and crossing test was obtained throughout this project.

6.2: Future direction/ plan

This project identified a number of putative causal mutations which have not yet been validated. A minor project is proposed to undertake the verification of the mutations in *Beyma*, WTS and WT. Specific primers will be designed to PCR-sequence the causal loci in order to validate if the SNPs are present only in *Beyma* and absent in WTS and WT. This analysis will be begun with analysing a mutation in an F-box family as mentioned earlier. Besides that, since the number of putative causal mutations was quite large; 126 in total, a few causal mutations will be selected as a start. If the causal mutations are an actual unique *Beyma* mutation, SNP segregation will be performed on F2 population of

cross between *Beyma* and Gifu with homozygous WT alleles. This step is required as a quick verification to determine if the SNPs are the ABA insensitive mutation, which will not be absent in F2 plants of homozygous WT carrier. Later, the sequence of the mutated locus will be stably transformed in to MG-20 WT using *Agrobacterium tumefaciens* transformation (Stiller et al., 1997; Lohar et al., 2001) to complement the homozygous recessive parent (reminder: the *Beyma* mutation shows dominant inheritance). Phenotype of the transformed WT will be analysed to verify the causative mutations of *Beyma*.

List of References

Asamizu E, Kato T, Sato S, Nakamura Y, Kaneko T and Tabata S (2003). Structural analysis of a *Lotus japonicus* genome. IV. Sequence features and mapping of seventy-three TAC clones which cover the 7.5 Mb regions of the genome. *DNA Research* 10: 115-122.

Austin RS, Vidaurre D, Stamatiou G, Breit R, Provart NJ, Bonetta D, Zhang J, Fung P, Gong Y, Wang PW, McCourt P and Guttman DS (2011). Next-generation mapping of *Arabidopsis* gene. *The Plant Journal* 67: 715-725.

Bano A and Harper JE (2002). Plant growth regulators and phloem exudates modulate root nodulation of soybean. *Functional Plant Biology* 29: 1299-1307.

Biswas B, Chan PC and Gresshoff PM (2009). A novel ABA insensitive mutant of *Lotus japonicus* with a wilty phenotype displays unaltered nodulation regulation. *Molecular Plant* 2: 487-499.

Bu Q, Lv T, Shen H, Luong P, Wang J, Wang Z, Huang Z, Xiao L, Engineer C, Kim TH, Schroeder JI and Huq E (2013). Regulation of drought tolerance by the F-box protein MAX2 in *Arabidopsis*. *Plant Physiology* 164: 424-439.

Caldwell DG, McCallum N, Shaw P, Muehlbauer GJ, Marshall DF and Waugh R (2004). A structured mutant population for forward and reverse genetics in barley (*Hordeum vulgare* L.). *The Plant Journal*. 40: 143-150.

Cannon SB, Crow JA, Heuer ML, Wang X, Cannon EKS, Dwan C, Lamblin AF, Vasdewani J, Mudge J, Cook A, Gish J, Cheung F, Kenton S, Kunau TM, Brown D, May GD, Kim D, Cook DR, Roe BA, Town CD, Young ND and Retzel EF (2005). Databases and information integration for the *Medicago truncatula* genome and transcriptome. *Plant Physiology* 138: 38-46.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Ruden DM and Lu X (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w^{1118}; *iso*-2; *iso*-3. *Landes Bioscience* 6: 1-13.

Cutler JA and Krochko JE (1999). Formation and breakdown of ABA. *Trends in Plant Science* 4: 472-478.

Cutler SR, Rodriguez PL, Finkelstein RR and Abrams SR (2010). Abscisic acid: emergence of a core signalling network. *Annu Rev Plant Biol* 61:651-679.

De Smet I, Zhang H, Inzé D and Beeckman T (2006). A novel role for abscisic acid emerges from underground. *Trends in Plant Science* 11: 434-439.

Ding Y, Kalo P, Yendrek C, Sun J, Liang , Marsh JF, Harris JM and Oldroyd GED (2008). Abscisic acid coordinates nod factor and cytokinin signalling during the regulation of nodulation in *Medicago truncatula*. *The Plant Cell* 20: 2681-2695.

Drummond AJ, Ashton B, Buxton S, Cheung M, Cooper A, Duran C, Field M, Heled J, Kearse M, Markowitz S, Moir R, Stones-Havas S, Sturrock S, Thierer T, Wilson A (2011). Geneious v5.4, available from http://www.geneious.com.

Edwards D, Batley J and Snowdon RJ (2013). Accessing complex crop genomes with next-generation sequencing. *Theoretical and Applied Genetics* 126: 1-11.

Ferguson BJ and Mathesius U (2003). Signalling interactions during nodule development. *Journal of Plant Growth Regulation* 22: 47-72.

Ferguson BJ, Indrasumunar A, Hayashi S, Lin MH, Lin YH, Reid DE and Gresshoff PM (2010). Molecular analysis of legume nodule development and autoregulation. *Journal of Integrative Plant Biology* 52: 61-76.

Fujii H and Zhu JK (2009). *Arabidopsis* mutant deficient in 3 abscisic acid-activated protein kinases reveals critical roles in growth, reproduction and stress. *Proceedings of the National Academy of Sciences* 106: 8380-8385.

Fujita M, Fujita Y, Noutoshi Y, Takahashi F, Narusaka Y, Yamaguchi-Shinozaki K and Shinozaki K (2006). Crosstalk between biotic and abiotic stress responses: a current view from the points of convergence in the stress signalling networks. *Current Opinion in Plant Biology* 9: 436-442.

Greene EA, Codomo CA, Taylor NE, Henikoff JG, Till BJ, Reynolds SH, Enns LC, Burtner C, Johnson JE, Odden AR, Comai L and Henikoff S (2003). Spectrum of chemically induced mutations from a large-scale reverse genetic screen in Arabidopsis. *Genetics.* 164: 731 – 740.

Handberg K and Stougaard J (1992). *Lotus japonicus*, an autogamous, diploid legume species for classical and molecular genetics. *The Plant Journal* 2: 487-496.

Hartwig B, James GV, Konrad K, Schneeberger K and Turck F (2012). Fast isogenic mapping-by-sequencing of ethyl methanesulfonate-induced mutant bulks. *Plant Physiology* 160: 591-600.

Hauser F, Waadt R and Schroeder JI (2011). Evolution of abscisic acid synthesis and signalling mechanisms. *Current Biology* 21: 346-355.

Hayashi M, Imaizumi-Anraku H, Akao S and Kawaguchi M (2000). Nodule organogenesis in *Lotus japonicus*. *Journal of Plant Research* 113: 489-495.

Hayashi M, Miyahara A, Sato S, Kato T, Yoshikawa M, Taketa M, Hayashi M, Pedrosa A, Onda R, Imaizumi-Anraku H, Bachmair A, Sandal N, Stougaard J, Murooka Y, Tabata S, Kawasaki S, Kawaguchi M and Harada K (2001). Construction of a genetic linkage map of the model legume *Lotus japonicus* using an intraspecific F_2 population. *DNA Research* 8: 301-310.

Hubbard KE, Nishimura N, Hitomi K, Getzoff ED and Schroeder JI (2010). Early abscisic acid signal transduction mechanisms: newly discovered components and newly emerging questions. *Genes and Development* 24: 1695-1708.

Jiang F and Hartung W (2008). Long-distance signalling of abscisic acid (ABA): the factor regulating the intensity of the ABA signal. *Journal of Experimental Botany* 59: 37-43.

Jiang Q and Gresshoff PM (1997). Classical and molecular genetics of the model legume *Lotus japonicus. Molecular Plant-Microbe Interactions* 10: 59-68.

Joshi-Saha A, Valon C and Leung J (2011). A brand new START: abscisic acid perception and transduction in the guard cell. *Science Signaling* 4: 1-13.

Kaneko T, Asamizu E, Kato T, Sato S, Nakamura Y and Tabata S (2003). Structural analysis of a *Lotus japonicus* genome. III. Sequence features and mapping of sixty-two TAC clones which cover the 6.7 Mb regions of the genome. *DNA Research* 10: 27-33.

Kato T, Kaneko T, Sato S, Nakamura Y and Tabata S (2000). Complete structure of the chloroplast genome of a legume, *Lotus japonicus*. *DNA Research* 7: 323-330.

Kato T, Sato S, Nakamura Y, Kaneko T, Asamizu E and Tabata S (2003). Structural analysis of a *Lotus japonicus* genome. V. Sequence features and mapping of sixty-four TAC clones which cover the 6.4 Mb regions of the genome. *DNA Research* 10: 277-285.

Kawaguchi M, Motomura T, Imaisumi-Anraku H, Akao S and Kawasaki S (2001). Providing the basis for genomics in *Lotus japonicus*: the accessions Miyakojima and Gifu are appropriate crossing partners for genetic analyses. *Molecular Genetics & Genomics* 266: 157-166.

Kawahara Y, de la BastideM, Hamilton JP, Kanamori H, McCombie WR, Ouyang S, Schwartz DC, Tanaka T, Wu J, Zhou S, Childs KL, Davidson RM, Lin H, Quesada-Ocampo L, Vaillancourt B, Sakai H, Lee SS, Kim J, Numa H, Itoh T, Buell CR, and Matsumoto T (2013). Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice* 6: 4.

Kazakoff SH, Imelfort M, Edwards D, Koehorst J, Biswas B, Batley J, Scott PT and Gresshoff PM (2012). Capturing the biofuel wellhead and powerhouse: the chloroplast and mitochondrial genomes of the leguminous feedstock tree *Pongamia pinnata*. PLoS ONE. 7: e51687.

Kermode AR (2005). Role of abscisic acid in seed dormancy. *Journal of Plant Regulation* 24: 319-344.

Kim JS, Mizoi J, Kidokoro S, Maruyama K, Nakajima J, Nakashima K, Mitsuda N, Takiguchi Y, Ohme-Takagi M, Kondou Y, Yoshizumi T, Matsui M, Shinozaki K and Yamaguchi-Shinozaki K (2012). *Arabidopsis* GROWTH-REGULATING FACTOR7 functions as a transcriptional repressor of abscisic acid- and osmotic stress-responsive genes, including *DREB2A*. *The Plant Cell* 24: 3393-3405.

Kim TH, Böhmer M, Hu H, Nishimura N and Schroeder JI (2010). Guard cell signal transduction network: advances in understanding abscisic acid, CO_2 and Ca^{2+} signalling. *Annual Review of Plant Biology* 61: 561-591.

Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK and Ding L (2009). VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* 25:2283-2285.

Koops P, Pelser S, Ignatz M, Klose C, Marrocco-Selden K and Kretsch T (2011). EDL3 is an F-box protein involved in the regulation of abscisic acid signaling in *Arabidopsis thaliana*. *Journal of Experimental Botany* 62: 5547-5560.

Koren S, Harhay GP, Smith TPL, Bono JL, Harhay DM, Mcvey SD, Radune D, Bergman NH and Phillippy AM (2013). Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biology* 14: R101.

Kuromori T, Sugimoto E and Shinozaki K (2014). Intertissue signal transfer of abscisic acid from vascular cells to guard cells. *Plant Physiology* 164: 1587-1592.

Lai CP, Lee CL, Chen PH, Wu SH, Yang CC and Shaw JF (2004). Molecular analyses of the *Arabidopsis* TUBBY-like protein gene family. *Plant Physiology* 134: 1586-1597.

Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, Karthikeyan AS, Lee CH, Nelson WD, Ploetz L, Singh S, Wensel A and Huala (2011). The Arabidopsis information resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research* 40: D1202-D1210.

Landrø NI (2014). Towards personalized treatment of depression: A candidate gene approach. *Scandinavian Journal of Psychology* 55: 219-224.

Lee KH, Piao HL, Kim HY, Choi SM, Jiang F, Hartung W, Hwang I, Kwak JM, Lee IJ and Hwang I (2006). Activation of glucosidase via stress-induced polymerisation rapidly increases active pools of abscisic acid. *Cell* 126: 1109-1120.

Li H, Sun J, Xu Y, Jiang H, Wu X and Li (2007). The bHLH-type transcription factor AtAIB positively regulates ABA response in *Arabidopsis*. *Plant Molecular Biology* 65: 655-665.

Li J, Dai X, Liu T and Zhao PX (2012). LegumeIP: an integrative database for comparative genomics and transcriptomics of model legumes. *Nucleic Acid Research* 40: D1221-D1229.

Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K and Wang J (2009). SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* 25: 1966-1967.

Liang Y and Harris JM (2005). Response of root branching to abscisic acid is correlated with nodule formation both in legumes and nonlegumes. *American Journal of Botany* 92: 1675-1683.

Liang Y, Mitchell DM and Harris JM (2007). Abscisic acid rescues the root meristem defects of the *Medicago truncatula latd* mutant. *Developmental Biology* 304: 297-307.

Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L and Law M (2012). Comparison of next-generaion sequencing systems. *Journal of Biomedicine and Biotechnology* 2012: Article ID 251364.

Liu Y, He J, Chen Z, Ren X, Hong X and Gong Z (2010). *ABA overly-sensitive 5* (*ABO5*), encoding a pentatricopeptide repeat protein required for *cis*-splicing of mitochondrial *nad2* intron 3, is involved in the abscisic acid response in *Arabidopsis*. *The Plant Journal* 63: 749-765.

Lohar DP, Schuller K, Buzas DM, Gresshoff PM and Stiller J (2001). Transformation of *Lotus japonicus* using the herbicide resistance *bar* gene as a selectable marker. *Journal of Experimental Botany* 52: 1697-1702.

Lorenc MT, Hayashi S, Stiller J, Lee H, Manoli S, Ruperao P, Visendi P, Berkman PJ, Lai K, Batley J and Edwards D (2012). Discovery of single nucleotide polymorphisms in complex genomes using SGSautoSNP. *Biology*. 1: 370-382.

Mardis ER (2007). The impact of next-generation sequencing technology on genetics. *Trends in Genetics* 24: 133-141.

Marshall DJ, Hayward A, Eales D, Imelfort M, Stiller J, Berkman PJ, Clark T, Mckenzie M, Lai K, Duran C, Batley and Edwards D (2010). Targeted identification of genomic regions using TAGdb. *Plant Methods* 6: 1-6.

McQuibban AG, Joza N, Megighian A, Scorzeto M, Zanini D, Reipert S, Richter C, Schweyen RJ and Nowikovsky K (2010). A *Drosophila* mutant of LETM1, a candidate gene for seizures in Wolf-Hirschhorn syndrome. *Human Molecular Genetics* 19: 987-1000.

Mei C, Jiang SC, Lu YF, Wu FQ, Yu YT, Liang S, Feng XJ, Comeras SP, Lu K, Wu Z, Wang XF and Zhang DP (2014). *Arabidopsis* pentatricopeptide repeat protein SOAR1 plays a critical role in abscisic acid signaling. *Journal of Experimental Botany* 65: 5317-5330.

Melchiorre M, Quero GE, Parola R, Racca R, Trippi VS and Lascano R (2009). Physiological characterization of four model Lotus diploid genotypes: *L. japonicus* (MG20 and Gifu), *L. filicaulis*, and *L. burttii* under salt stress. *Plant Science* 177: 618-628.

Merlot S, Gosti F, Guerrier D, Vavasseur A and Giraudat J (2001). The ABI1 and ABI2 protein phosphatises 2C act in a negative feedback regulatory loop of the abscisic acid signalling pathway. *The Plant Journal* 25: 295-303.

Metzker ML (2010). Sequencing technologies – the next generation (2010). *Nature Reviews Genetics* 11: 31-46.

Miao Y, Lv D, Wang P, Wang XC, Chen J, Miao C and Song CP (2006). An *Arabidopsis* glutathione peroxidase functions as both a redox transducer and a scavenger in abscisic acid and drought stress responses. *The Plant Cell* 18: 2749-2766.

Minoia S, Petrozza A, D'Onofrio O, Piron F, Mosca G, Sozio G, Cellini F, Bendahmane A and Carriero F (2010). A new mutant genetic resource for tomato crop improvement by TILLING technology. *BMC Research Notes*. 3: 69.

Nakamura Y, Kaneko T, Asamizu E, Kato T, Sato S and Tabata S (2002). Structural analysis of a *Lotus japonicus* genome. II. Sequence features and mapping of sixty-five TAC clones which cover the 6.5-Mb regions of the genome. *DNA Research* 9: 63-70.

Nakashima K and Yamaguchi-Shinozaki K (2013). ABA signaling in stress-response and seed development. *Plant Cell Reports* 32: 959-970.

Nambara E and Marion-Poll A (2005). Abscisic acid biosynthesis and catabolism. *Annual Review of Plant Biology* 56: 165-185.

Ng LM, Melcher K, Teh BT and Xu HE (2014). Abscisic acid perception and signaling: structural mechanisms and application. *Acta Pharmacologica Sinica* 35: 567-584.

Nordström KJV, Albani MC, James GV, Gutjahr C, Hartwig B, Turck F, Paszkowski U, Coupland G and Schneeberger K (2013). Mutation identification by direct comparison of whole-genome sequencing data from mutant and wild-type individuals using *k*-mers. *Nature Biotechnology* 31: 325-331.

Okamoto M, Tanaka Y, Abrams SR, Kamiya Y, Seki M and Nambara E (2009). *High humidity induces abscisic acid 8'-hydroxylase in stomata and vasculature to regulate local and systemic abscisic acid responses in Arabidopsis. Plant Physiology 149: 825-834.*

Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova M, Krabichler B, Speicher MR, Zschocke J and Trajanoski Z (2013). A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in Bioinformatics* doi:10.1093/bib/bbs086.

Pareek CS, Smoczynski R and Tretyn A (2011). Sequencing technologies and genome sequencing. *Journal of Applied Genetics* 52: 413-435.

Patel S and Patel NK (2013). Candidate gene mapping: approach, methods and significance. *American Journal of Research Communication* 1: 199-204.

Perry J, Brachmann A, Welham T, Binder A, Charpentier M, Groth M, Haage K, Markmann K, Wang TL and Parniske M (2009). TILLING in *Lotus japonicus* identified large allelic series for symbiosis genes and revealed a bias in functionally defective ethyl methanesulfonate alleles toward glycine replacements. *Plant Physiology*. 151: 1281-1291.

Pflieger S, Lefebvre V and Causse M (2001). The candidate gene approach in plant genetics: a review. *Molecular Breeding* 7: 275-291.

Piertney SB, Webster LMI (2010). Characterising functionally important and ecologically meaningful genetic diversity using a candidate gene approach. *Genetica* 138: 419-432.

Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP and Gu Y (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* 13: 341.

Raghavendra AS, Gonugunta VK, Christmann A and Grill E (2010). ABA perception and signalling. *Trends in Plant Science* 15: 395-401.

Rock CD, Sakata Yand Quatrano RS (2010). Stress signaling I: The role of abscisic acid (ABA). In *Abiotic Stress Adaptation in Plants: Physiological, Molecular and Genomic Foundation.* Edited by Pareek A, Sopory SK, Bohnert HJ and Govindjee. Dortrecht: Springer, pg 33-73.

Saeki K and Kouchi H (2000). The Lotus symbiont, *Mesorhizobium loti*: Molecular genetic techniques and application. *Journal of Plant Research* 113: 457-465.

Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes JC, Hutchison III CA, Slocombe PM and Smith M (1977b). Nucleotide sequence of bacteriophage ΦX174 DNA. *Nature* 265: 687-695.

Sanger F, Nicklen S and Coulson AR (1977a). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States* 74: 5463-5467.

Sato S and Andersen SU (2014). Genome sequencing. *In* The Lotus japonicus Genome, Compendium of Plant Genomes. Tabata S and Stougaard J (eds.) Springer-Verlag Berlin Heidelberg: 35-40.

Sato S and Tabata S (2005). *Lotus japonicus* as a platform for legume research. *Current Opinion in Plant Biology* 9: 128-132.

Sato S, Kaneko T, Nakamura Y, Asamizu E, Kato T and Tabata S (2001). Structural analysis of a *Lotus japonicus* genome. I. Sequence features and mapping of fifty-six TAC clones which cover the 5.4 Mb regions of the genome. *DNA Research* 8: 311-318.

Sato S, Nakamura Y, Asamizu E, Isobe S and Tabata S (2007). Genome sequencing and genome resources in model legumes. *Plant Physiology* 144: 588-593.

Sato S, Nakamura Y, Kaneko T, Asamizu E, Kato T, Nakao M, Sasamoto S, Watanabe A, Ono A, Kawashima K, Tsunakazu F, Katoh M, Kohara M, Kishida Y, Minami C, Nakayama S, Nakazaki N, Shimizu Y, Shinpo S, Takahashi C, Wada T, Yamada M, Ohmido N, Hayashi M, Fukui K, Baba T, Nakamichi T, Mori H and Tabata S (2008). Genome structure of the legume, *Lotus japonicus*. *DNA Research* 15: 227-239.

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, Xu D, Hellsten U, May GD, Valliyodan B, Lindquist E, Peto M, Grant D, Shu S, Goodstein D, Barry K, Futrell-Griggs M, Abernathy B, Du J, Tian Z, Zhu L, Gill N, Joshi T, Libault M, Sethuraman A, Zhang XC, Shinozaki K, Nguyen HT, Wing RA, Cregan P, Specht J, Grimwood J, Rokhsar D, Stacey G, Shoemaker RC and Jackson SA (2010). Genome sequence of the palaeopolyploid soybean. *Nature* 463: 178-183.

Schmutz J, McClean PE, Mamidi S, Wu GA, Cannon SB, Grimwood J, Jenkins J, Shu S, Song Q, Chavarro C, Torres-Torres M, Geffroy V, Moghaddam SM, Gao D, Abernathy B, Barry K, Blair M, Brick MA, Chovatia M, Gepts P, Goodstein DM, Gonzales M, Hellsten U, Hyten DL, Jia G, Kelly JD, Kudrna D, Lee R, Richard MMS, Miklas PN, Osorno JM, Rodrigues J, Thareau V, Urrea CA, Wang M, Yu Y, Zhang M, Wing RA, Cregan PB, Rokhsar DS and Jackson SA (2014). A reference genome for common bean and genome-wide analysis of dual domestications. *Nature Genetics* 46: 707-713.

Schroeder JI (1992). Plasma membrane ion channel regulation during abscisic acid-induced closing of stomata. *Philosophical Transactions: Biological Sciences* 338: 83-89.

Shan H, Chen S, Jiang J, Chen F, Chen Y, Gu C, Li P, Song A, Zhu X, Gao H, Zhou G, Li T and Yang X (2011). Heterologous expression of the chrysanthemum R2R3-MYB transcription factor *CmMYB2* enhances drought and salinity tolerance, increases

hypersensitivity to ABA and delays flowering in *Arabidopsis thaliana*. *Molecular Biotechnology* 51: 160-173.

Shiwa Y, Fukushima-Tanaka S, Kasahara K, Horiuchi T and Yoshikawa H (2012). Whole-genome profiling of a novel mutagenesis technique using proofreading-deficient DNA polymerase δ. *International Journal of Evolutionary Biology*. DOI: 10.1155/2012/860797.

Sims D, Sudbery I, Ilott NE, Heger A and Ponting CP (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics* 15: 121-132.

Sirichandra C, Wasilewska A, Vlad F, Valon C and Leung J (2009). The guard cell as a single-cell model towards understanding drought tolerance and abscisic acid action. *Journal of Experimental Botany* 60: 1439-1463.

Smadja CM, Canbäck B, Vitalis R, Gautier M, Ferrari J, Zhou JJ and Butlin RK (2012). Large-scale candidate gene scan reveals the role of chemoreceptor genes in host plant specialisation and speciation in the pea aphid. *Evolution* 66: 2723-2738.

Snustad P and Simmons MJ (2003). Principles of genetics: Third edition. John Wiley & Sons, Inc, New Jersey.

Stacey G, Libault M, Brechenmacher L, Wan J and May GD (2006). Genetics and functional genomics of legume nodulation. *Current Opinion in Plant Biology* 9: 110-121.

Steffens B, Wang J and Sauter M (2006). Interactions between ethylene, gibberellins and abscisic acid regulate emergence and growth rate of adventitious roots in deepwater rice. *Planta* 223: 604-612.

Stiller J, Martirani L, Tuppale S, Chian RJ, Chiurazzi M and Gresshof PM (1997). High frequency transformation and regeneration of transgenic plants in the model legume *Lotus japonicus*. *Journal of Experimental Botany* 48: 1357-1365.

Suzuki A, Akune M, Kogiso M, Imagama Y, Osuki K, Uchiumi T, Higashi S, Han SY, Yoshida S, Asami T and Abe M (2004). Control of nodule number by the phytohormone abscisic acid in the roots of two leguminous species. *Plant Cell Physiology* 45: 914-922.

Szczyglowski K and Stougaard J (2008). *Lotus* genome: pod of gold for legume research. *Trends in Plant Science* 13: 515-517.

Tabata R, Kamiya T, Shigenobu S, Yamaguchi K, Yamada M, Hasebe M, Fujiwara T, Sawa S (2013). Identification of an EMS-induced causal mutation in a gene required for boron-mediated root development by low coverage genome re-sequencing in *Arabidopsis*. *Plant Signal & Behavior* 8: e22534.

Tan J, Tan Z, Wu F, Sheng P, Heng Y, Wang X, Ren Y, Wang J, Guo X, Zhang X, Cheng Z, Jiang L, Liu X, Wang H and Wan J (2014). A novel chloroplast-localised

pentratricopeptide repeat protein involved in splicing affects chloroplast development and abiotic stress response in rice. *Molecular Plant* 7: 1329-1349.

The Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408: 796-815.

Thudi M, Li Y, Jackson SA, May GD and Varshney RK (2012). Current state-of-art of sequencing technologies for plant genomics research. *Briefing in Functional Genomics*. 11: 3-11.

Till BJ, Cooper J, Ti TH, Colowit P, Greene EA, Henikoff S and Comai L (2007). Discovery of chemically induced mutations in rice by TILLING. *BMC Plant Biology*. 7: 19.

Tominaga A, Nagata M, Futsuki K, Abe H, Uchiumi T, Abe M, Kucho K, Hashiguchi M, Akashi R, Hirsch A, Arima S and Suzuki A (2010). Effect of abscisic acid on symbiotic nitrogen fixation activity in the root nodules of *Lotus japonicus*. *Plant Signalling and Behavior* 5: 440-443.

Uchida N, Sakamoto T, Kurata T and Tasaka M (2011). Identification of EMS-induced causal mutations in a non-reference *Arabidopsis thaliana* accession by whole genome sequencing. *Plant and Cell Physiology* 52: 716-722.

Udvardi MK (2001). Legume models strut their stuff. *Molecular Plant-Microbe Interactions* 14: 6-9.

Udvardi MK, Tabata S, Parniske M and Stougaard J (2005). *Lotus japonicus*: legume research in the fast lane. *Trends in Plant Science* 10: 222-228.

Umezawa T, Nakashima K, Miyakawa T, Kuromori T, Tanokura M, Shinozaki K and Yamaguchi-Shinozaki K (2010). Molecular basis of the core regulatory network in ABA responses: sensing, signalling and transport. *Plant and Cell Physiology* 51: 1821-1839.

Varshney RK, Nayak SN, May GD and Jackson SA (2009). Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends in Biotechnology* 27: 522-530.

Varshney RK, Terauchi R and McCouch SR (2014). Harvesting the promising fruits of genomics: Applying genome sequencing technologies to crop breeding. *PLOS Biology* 12: e1001883.

Wang RS, Pandey S, Li S, Gookin TE, Zhao Z, A R and Assmann SM (2011). Common and unique elements of the ABA-regulated transcriptome of Arabidopsis guard cells. *BMC Genomics* 12: 216.

Wasilewska A, Vlad F, Sirichandra C, Redko Y, Jammes F, Valon C, Frey NF and Leung J (2008). An update on abscisic acid signalling in plants and more... *Molecular Plant* 1: 198-217.

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song X, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA and Rothberg JM (2008). The complete genome of an individual by massively paralled DNA sequencing. *Nature* 452: 872-877.

Yang C, Li A, Zhao Y, Zhang Z, Zhu Y, Tan X, Geng S, Guo H, Zhang X, Kang Z and Mao L (2011). Overexpression of a wheat CCaMK gene reduces ABA sensitivity of *Arabidopsis thaliana* during seed germination and seedling growth. *Plant Molecular Biology Reporter* 29: 681-692.

Yendrek CR, Lee YC, Morris V, Liang Y, Pislariu CI, Burkart G, Meckfessel MH, Salehin M, Kessler H, Wessler H, Lloyd M, Lutton H, Teillet A, Sherrier DJ, Journet EP, Harris JM and Dickstein R (2010). A putative transporter is essential for integrating nutrient and hormone signalling with lateral root growth and nodule development in *Medicago truncatula*. *The Plant Journal* 62: 100-112.

Yoshida T, Fujita Y, Maruyama K, Mogami J, Todaka D, Shinozaki K and Yamaguchi-Shinozaki K (2014). Four *Arabidopsis* AREB/ABF transcription factors function predominantly in gene expression downstream of SnRK2 kinases in abscisic-acid signaling in response to osmotic stress. *In press.*

Young ND, Cannon SB, Sato S, Kim D, Cook DR, Town CD, Roe BA and Tabata S (2005). Sequencing the genespaces of *Medicago truncatula* and *Lotus japonicus*. *Plant Physiology* 137: 1174-1181.

Yu X and Sun S (2013). Comparing a few SNP calling algorithms using low-coverage sequencing data. *BMCB Bioinformatics* 14: 274.

Zander M, Patel DA, Van de Wouw A, Lai K, Lorenc MT, Campbell E, Hayward A, Edwards D, Raman H and Batley J (2013). Identifying genetic diversity of avirulence genes in *Leptosphaeria maculans* using whole genome sequencing. *Functional and Integrative Genomics* 13: 295-308.

Zhang W, Chen J, Yang Y, Tang Y, Shang J and Shen B (2011). A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies. *PLOS ONE* doi: 10.1371/journal.pone.0017915

Zhang Y and Xue Y (2009). DOR: a link between an F-box protein and guard cell ABA signaling. *Plant Signaling & Behavior* 4: 470-471.

Zhang Y, Xu W, Li Z, Deng XW, Wu W and Xue Y (2008). F-box protein DOR functions as a novel inhibitory factor for abscisic acid-induced stomatal closure under drought stress in *Arabidopsis*. *Plant Physiology* 148: 2121-2133.

Zhu Q, Smith SM, Ayele M, Yang L, Jogi A, Chaluvadi SR and Bennetzen JL (2012a). High-throughput discovery of mutations in tef semi-dwarfing genes by next-generation sequencing analysis. *Genetics* 129: 819-829.

Zhu Y, Mang H, Sun Q, Hipps A and Hua J (2012b). Gene discovery using mutagen-induced polymorphisms and deep sequencing: application to plant disease resistance. *Genetics* 192: 139-146.

Zuryn S, Gras SL, Jamet K and Jarriault (2010). A strategy for direct mapping and identification of mutations by whole-genome sequencing. *Genetics* 186: 427-430.

Appendices

Table A1: Details on putative causal SNPs in *Beyma* from the sequencing of single genomes.

Chro	Position	Ref	Change	Quality	CV	Gene name	Effect	AA change	Codon change	Codon number	Putative function
0	12184744	A	G	2	13	LjT45D24.80.r2.d	NS	H/R	cAc/cGc	534	WRKY transcription factor 43
1	17947055	T	G	2	4	chr1.LjT10C23.10.r2.m	NS	Q/P	cAg/cCg	419	Transposase
1	21924395	T	C	2	7	chr1.LjT39K06.170.r2.d	NS	M/T	aTg/aCg	122	Serine protease inhibitor (SERPIN) family protein
1	36310579	A	C	2	10	chr1.LjT11G20.70.r2.d	NS	T/P	Acc/Ccc	6	tRNA guanosine-2'-O-methyltransferase
1	48316127	A	C	2	6	chr1.CM0104.730.r2.a	NS	V/G	gTg/gGg	116	Histone-lysine N-methyltransferase
1	55718667	T	G	2	19	chr1.CM0322.120.r2.a	NS	F/C	tTt/tGt	257	Tetrahydrofolate dehydrogenase/cyclohydrolase
1	57062066	A	G	2	11	chr1.CM0295.900.r2.m	NS	Y/H	Tac/Cac	128	Unknown protein
1	64364378	T	C	2	19	chr1.CM0544.330.r2.m	NS	M/V	Atg/Gtg	445	Pentatricopeptide repeat-containing protein
2	151028	A	G	2	9	chr2.CM0067.290.r2.m	NS	E/G	gAg/gGg	425	Cytochrome P450 monooxygenase
2	4957714	A	G	2	11	chr2.LjT24G10.90.r2.d	NS	V/A	gTa/gCa	89	Unknown protein
2	7951761	A	C	2	18	chr2.LjT16L14.40.r2.m	NS	K/Q	Aag/Cag	156	Unknown protein
2	9228954	G	T	2	17	chr2.CM0435.950.r2.a	NS	A/D	gCc/gAc	534	Homeobox-leucine zipper protein
2	18758656	C	G	2	6	chr2.CM0124.100.r2.d	NS	D/H	Gat/Cat	125	ATP binding protein
2	21311265	C	G	2	15	chr2.CM0608.1080.r2.d	NS	W/S	tGg/tCg	110	Cystathionine beta-synthase (CBS) protein
2	27033357	C	G	2	14	chr2.CM0249.380.r2.d	NS	P/A	Ccg/Gcg	132	Leucine-rich receptor-like protein kinase
2	27277434	A	C	2	13	chr2.CM0249.840.r2.m	NS	H/P	cAt/cCt	547	Aldehyde dehydrogenase
2	31511421	T	G	2	8	chr2.CM0021.3410.r2.m	NS	E/D	gaA/gaC	5	DNA-binding PD1-like protein
2	44016442	T	G	2	29	chr2.CM0191.880.r2.m	NS	I/L	Att/Ctt	550	Endomembrane protein 70 protein family
3	4145569	T	G	2	10	chr3.CM0282.950.r2.m	NS	E/A	gAa/gCa	53	UDP-Glycosyltransferase superfamily protein

3	5066525	T	C	2	8	chr3.LjT13N17.110.r2.m	NS	V/A	gTg/gCg	206	RING/U-box superfamily protein
3	6464930	A	G	2	23	chr3.CM0574.220.r2.m	NS	F/L	Ttt/Ctt	210	Unknown protein
3	10831078	C	T	2	6	chr3.CM0451.1060.r2.d	NS	E/K	Gag/Aag	240	F-box family protein
3	16032774	T	C	2	10	chr3.LjT33C23.90.r2.d	NS	E/G	gAg/gGg	197	NAC domain protein
3	20316472	T	C	2	22	chr3.LjT14J20.20.r2.m	NS	V/A	gTt/gCt	274	Unknown protein
3	22723971	G	C	2	16	chr3.CM0619.120.r2.d	NS	V/L	Gta/Cta	71	Anthocyanidin 3-O-glucosyltransferase 2
3	30256701	T	C	2	6	chr3.CM0226.190.r2.m	NS	N/S	aAc/aGc	226	Gibberellin 3-beta-hydroxylase
3	31755293	T	G	2	18	chr3.CM0213.690.r2.m	NS	Q/P	cAg/cCg	273	Nuclear poly(a) polymerase
3	33745046	A	C	2	4	chr3.CM0208.10.r2.d	NS	L/V	Ttg/Gtg	589	Subtilisin-like serine protease
3	39899374	A	C	2	9	chr3.CM0164.20.r2.d	NS	K/T	aAg/aCg	72	ABA Deficient 2
3	41967562	C	G	2	13	chr3.CM0616.110.r2.d	NS	Q/E	Cag/Gag	27	MAP kinase 4
4	3238390	A	G	2	17	chr4.CM0007.920.r2.d	NS	V/A	gTc/gCc	106	Minichromosome maintenance (MCM2/3/5) family protein
4	5211180	A	G	2	12	chr4.CM0100.410.r2.d	NS	N/S	aAc/aGc	11	Unknown protein
4	10163451	T	G	2	14	chr4.CM0227.170.r2.d	NS	K/Q	Aaa/Caa	17	Fucosyltransferase 12
4	21905007	A	G	2	12	chr4.CM0126.2110.r2.a	NS	H/R	cAc/cGc	24	Dehydration responsive element binding protein 1
4	23321764	T	G	2	9	chr4.CM0173.140.r2.m	NS	C/G	Tgc/Ggc	257	BTB/POZ domain-containing protein
4	25280700	T	C	2	14	chr4.CM0087.130.r2.m	NS	H/R	cAc/cGc	278	Geranylgeranyl reductase
4	32266662	T	C	2	11	chr4.CM0006.530.r2.m	NS	N/D	Aac/Gac	307	Acyl-transferase family protein
4	40092456	A	G	2	13	chr4.CM0004.1210.r2.m	NS	V/A	gTc/gCc	28	Unknown protein
4	42766122	T	G	2	16	chr4.CM0042.1870.r2.m	NS	N/H	Aat/Cat	11	poly(A) polymerase 1
5	2688010	T	C	2	8	chr5.CM0852.160.r2.m	NS	Q/R	cAg/cGg	14	Abscisic acid 8'-hydroxylase
5	2958313	T	G	2	12	chr5.CM0096.20.r2.m	NS	V/G	gTt/gGt	322	L-asparaginase
5	3441906	T	G	2	26	chr5.CM0096.900.r2.d	NS	L/R	cTc/cGc	419	Exocyst complex component 7
5	5846695	T	C	2	9	chr5.CM0345.240.r2.d	NS	D/G	gAc/gGc	27	Tudor/PWWP/MBT superfamily protein
5	13034937	T	C	2	13	chr5.CM0300.90.r2.d	NS	H/R	cAc/cGc	349	Pentatricopeptide repeat (PPR) superfamily protein
5	13880165	A	G	2	11	chr5.CM0571.190.r2.a	NS	Y/C	tAc/tGc	56	Unknown protein
5	13911038	T	C	2	5	chr5.CM0571.250.r2.m	NS	V/A	gTg/gCg	22	Syntaxin-121

5	26269990	T	G	2	12	chr5.CM0239.480.r2.m	NS	Y/D	Tat/Gat	161	Heat shock protein DnaJ with tetratricopeptide repeat
5	32604650	A	C	2	7	chr5.CM0200.1280.r2.d	NS	L/R	cTt/cGt	123	Copper amine oxidase
5	35834052	T	G	2	15	chr5.CM1439.220.r2.d	NS	E/A	gAg/gCg	79	Clathrin adaptor complexes medium subunit family protein
5	36283759	T	G	2	20	chr5.CM0180.270.r2.m	NS	S/A	Tct/Gct	95	2-oxoglutarate dehydrogenase
6	844905	A	G	2	16	chr6.CM1613.310.r2.m	NS	E/G	gAg/gGg	40	TRAF-like superfamily protein
6	8201050	C	A	2	8	chr6.LjT111F18.50.r2.m	NS	P/H	cCt/cAt	524	Unknown protein
6	18949733	T	G	3	7	chr6.LjT35H04.120.r2.d	NS	F/V	Ttt/Gtt	102	Central motor kinesin 1
6	20259867	T	C	2	20	chr6.CM0437.210.r2.m	NS	L/P	cTt/cCt	238	Reticuline oxidase
6	20344197	G	C	2	19	chr6.CM0437.400.r2.m	NS	L/V	Ctc/Gtc	471	Unknown protein
6	22420967	A	C	2	5	chr6.CM0139.460.r2.m	NS	S/A	Tcg/Gcg	311	Aspartyl protease family protein
6	25194091	T	G	2	15	chr6.CM0114.200.r2.m	NS	K/T	aAa/aCa	7	Unknown protein
4	22771513	A	G	2	20	chr4.CM1864.490.r2.a	SSA				Bromodomain-containing factor 1
6	14026407	T	G	2	19	chr6.CM0037.160.r2.m	SSD				WD-40 repeat family protein

Chro: chromosome; Ref: reference base; CV: coverage; AA: amino acid; NS: nonsynonymous change; SSA: splice site acceptor; SSD: splice site donor.

Table A2: Details on putative causal SNPs in *Beyma* from the re-sequencing of pooled genomes.

Chro	Position	Ref	Change	Quality	CV	Gene name	Effect	AA change	Codon change	Codon Number	Putative function
1	438983	G	T	2	19	chr1.CM0088.930.r2.m	NS	A/E	gCg/gAg	176	Gibberellin 2-oxidase 8
1	16199564	G	A	2	12	chr1.CM0320.470.r2.m	NS	V/I	Gtt/Att	610	Pentatricopeptide repeat superfamily protein
1	25846212	T	A	2	13	chr1.CM0442.510.r2.d	NS	H/Q	caT/caA	172	AGAMOUS-like 92
1	26036677	C	A	2	11	chr1.CM0760.150.r2.d	NS	D/Y	Gat/Tat	205	Unknown protein
1	34267949	A	G	2	8	chr1.CM0393.260.r2.d	NS	M/T	aTg/aCg	676	AP2-like ethylene-responsive transcription factor
1	35012700	C	T	2	4	chr1.LjT29L18.90.r2.d	NS	G/D	gGt/gAt	140	Hyaluronan / mRNA binding family
1	35305238	C	T	4	14	chr1.CM0051.230.r2.m	NS	A/V	gCa/gTa	555	Phosphoglycerate mutase-like family protein
1	48422659	C	A	2	12	chr1.CM0104.800.r2.d	NS	V/L	Gta/Tta	25	Receptor-like protein kinase 2
1	49856234	G	A	3	12	chr1.CM0104.2750.r2.a	NS	A/T	Gcg/Acg	137	Anthranilate N-benzoyltransferase protein 2
1	60885565	C	A	2	6	chr1.CM0029.580.r2.d	SG	E/*	Gag/Tag	136	F-box family protein
1	64353003	G	C	2	12	chr1.CM0544.300.r2.m	NS	R/T	aGa/aCa	655	Mitogen-activated protein kinase
1	65798280	G	T	2	16	chr1.CM0105.670.r2.a	NS	R/I	aGa/aTa	414	Unknown protein
1	65862187	G	A	5	20	chr1.CM0105.760.r2.a	NS	A/T	Gct/Act	469	Alpha/beta-Hydrolases superfamily protein
1	65978176	G	T	2	8	chr1.CM0105.920.r2.m	NS	L/M	Ctg/Atg	322	Aspartyl protease family protein
1	66040218	G	T	2	14	chr1.CM0105.1020.r2.m	NS	Q/H	caG/caT	422	Auxin response factor 9
2	7968180	C	T	8	32	chr2.LjT16L14.60.r2.m	NS	G/D	gGc/gAc	113	Unknown protein
2	8558599	G	A	2	8	chr2.CM0435.230.r2.d	NS	S/L	tCa/tTa	704	Disease resistance protein
2	20948830	T	C	2	11	chr2.CM0608.680.r2.d	NS	E/G	gAg/gGg	347	Methyltransferases superfamily protein
2	24577588	G	T	2	15	chr2.CM0230.110.r2.m	NS	A/S	Gct/Tct	480	Cellulose synthase-like protein

2	24820523	C	T	2	17	chr2.CM0020.130.r2.d	NS	D/N	Gat/Aat	291	Unknown protein
2	26186218	A	G	2	17	chr2.CM0272.170.r2.m	NS	N/S	aAc/aGc	200	Phosphoserine aminotransferase
2	27302037	G	C	2	15	chr2.CM0249.870.r2.m	NS	C/S	tGc/tCc	633	Methyltransferases superfamily protein
2	27430246	A	T	2	8	chr2.CM0249.1180.r2.d	SG	C/*	tgT/tgA	305	Unknown protein
2	29476266	C	A	2	9	chr2.CM0021.270.r2.m	NS	L/F	ttG/ttT	946	Receptor protein kinase-like protein
2	31484615	G	A	4	14	chr2.CM0021.3350.r2.m	NS	S/F	tCt/tTt	86	Glycosyltransferase family protein 28
2	39047845	C	T	5	12	chr2.LjT43K05.170.r2.d	NS	T/I	aCt/aTt	189	RNA-binding KH domain-containing protein
2	44270569	G	T	2	13	chr2.CM0102.250.r2.m	NS	Q/H	caG/caT	457	Nodule inception protein
2	44369531	C	A	2	11	chr2.CM0102.440.r2.d	NS	P/T	Cca/Aca	103	Glutamate receptor 2
3	871417	C	T	3	21	chr3.LjT34H24.170.r2.d	NS	V/I	Gtt/Att	107	Nucleoporin interacting component family protein
3	1827782	C	A	2	7	chr3.CM1488.480.r2.a	SG	E/*	Gag/Tag	445	Unknown protein
3	3917400	G	A	8	16	chr3.CM0282.610.r2.m	NS	S/F	tCt/tTt	960	ATPase family AAA domain-containing protein 2B
3	9215269	C	A	2	14	chr3.LjT47H21.70.r2.a	NS	C/F	tGc/tTc	149	Cysteine desulfurase
3	10831078	C	T	4	11	chr3.CM0451.1060.r2.d	NS	E/K	Gag/Aag	240	F-box family protein
3	12873676	G	A	2	11	chr3.CM0279.620.r2.m	SG	R/*	Cga/Tga	225	Pentatricopeptide repeat superfamily protein
3	16383602	A	C	2	5	chr3.CM0196.10.r2.a	NS	Y/D	Tat/Gat	93	Serine/threonine-protein phosphatase
3	33564606	C	A	2	25	chr3.LjT07B06.40.r2.m	NS	R/M	aGg/aTg	121	Pyridoxamine 5'-phosphate oxidase family protein
3	34272982	G	T	2	17	chr3.CM0416.110.r2.m	NS	V/L	Gtg/Ttg	151	Unknown protein
3	34317759	G	A	2	10	chr3.CM0416.180.r2.m	NS	A/T	Gca/Aca	18	Sugar transporter superfamily
3	34612898	T	A	2	23	chr3.CM0416.630.r2.a	SG	C/*	tgT/tgA	76	SAP domain-containing protein
3	35538004	C	A	2	10	chr3.CM0115.150.r2.d	NS	S/Y	tCt/tAt	17	Unknown protein
3	36187033	C	A	2	11	chr3.CM0049.620.r2.m	NS	M/I	atG/atT	152	Pyridoxal phosphate phosphatase
3	40848213	G	A	5	15	chr3.CM0396.380.r2.d	NS	S/N	aGt/aAt	37	Unknown protein

3	43473057	G	A	4	9	chr3.CM0091.410.r2.m	NS	S/F	tCc/tTc	185	F-box and wd40 domain protein
3	44334061	C	T	2	13	chr3.CM0091.1700.r2.m	NS	S/F	tCt/tTt	669	S-locus lectin protein kinase family protein
3	44410057	C	A	2	13	chr3.CM0091.1730.r2.m	NS	G/V	gGt/gTt	84	Kinesin motor family protein
3	46128748	C	T	3	9	chr3.CM0460.180.r2.d	NS	P/L	cCc/cTc	242	Unknown protein
4	553123	G	A	2	9	chr4.CM0525.170.r2.m	NS	R/H	cGc/cAc	86	Phosphofructokinase 2
4	1097367	C	T	3	11	chr4.CM0288.670.r2.m	SG	Q/*	Caa/Taa	191	RHO protein GDP dissociation inhibitor
4	1249844	G	A	2	12	chr4.CM0288.930.r2.d	NS	A/T	Gca/Aca	202	Plant UBX domain containing protein 4
4	2005542	G	A	3	14	chr4.CM0026.910.r2.m	NS	R/H	cGc/cAc	336	Glutamate decarboxylase
4	2827197	G	T	2	13	chr4.CM0007.340.r2.m	NS	L/F	ttG/ttT	313	FtsH extracellular protease family
4	4276531	T	G	2	9	chr4.CM0337.800.r2.m	NS	T/P	Acg/Ccg	239	Amine oxidase
4	10476144	C	A	2	17	chr4.CM0227.530.r2.m	NS	D/Y	Gac/Tac	172	Filament-like plant protein 7
4	11670000	G	T	2	5	chr4.CM0075.10.r2.m	NS	P/T	Ccc/Acc	60	Pentatricopeptide repeat-containing protein A
4	12680132	C	T	4	13	chr4.CM0128.600.r2.m	NS	L/F	Ctc/Ttc	166	Unknown protein
4	12830788	C	A	2	4	chr4.CM0128.800.r2.m	NS	A/E	gCa/gAa	86	Ribonuclease III family protein
4	12836641	C	T	9	28	chr4.CM0128.810.r2.m	NS	L/F	Ctt/Ttt	135	Casein kinase alpha 1
4	14962530	C	A	2	13	chr4.CM0165.710.r2.d	NS	A/S	Gca/Tca	204	Succinate dehydrogenase
4	18677529	G	T	2	4	chr4.CM0161.220.r2.d	NS	D/Y	Gat/Tat	45	Cysteine-rich receptor-like protein kinase 2
4	32262978	G	A	2	16	chr4.CM0006.520.r2.m	NS	V/I	Gtt/Att	1734	Alpha/beta-Hydrolases superfamily protein
4	39456672	T	C	2	15	chr4.CM0004.390.r2.d	NS	K/R	aAg/aGg	150	Bromo-adjacent homology domain-containing protein
4	40895695	G	T	2	9	chr4.CM0004.2380.r2.d	NS	F/L	ttC/ttA	76	CCCH-type zinc finger family protein
4	41611281	A	G	2	10	chr4.CM0042.70.r2.d	NS	I/T	aTa/aCa	104	Unknown protein
4	43171368	C	T	4	13	chr4.CM0042.2570.r2.d	NS	A/V	gCc/gTc	393	ABC-2 type transporter family protein

5	3248015	G	A	5	23	chr5.CM0096.510.r2.a	NS	S/N	aGc/aAc	231	Transcription factor bHLH85
5	5878147	C	T	4	10	chr5.CM0345.300.r2.d	NS	P/L	cCa/cTa	165	ARM repeat superfamily protein
5	5999682	G	A	2	12	chr5.CM0345.460.r2.m	NS	P/L	cCt/cTt	1	Mitochondrial glycoprotein family protein
5	36193530	G	A	3	14	chr5.CM0180.70.r2.d	NS	R/K	aGg/aAg	64	Leucine-rich repeat (LRR) family protein
6	24544529	C	T	2	18	chr6.CM0118.930.r2.d	NS	E/K	Gag/Aag	312	AAA-type ATPase family protein

Chro: chromosome; Ref: reference base; CV: coverage; AA: amino acid; NS: nonsynonymous change; SG: stop gained change.